LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

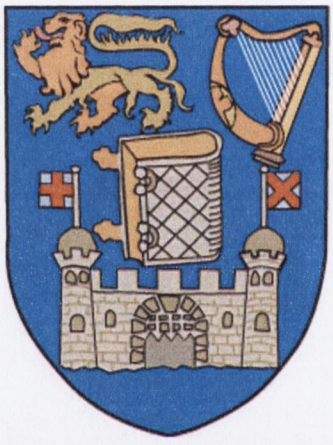
Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

GENETIC VARIATION IN THE BOVINE GENOME: SIGNATURES OF SELECTION AND DEMOGRAPHIC HISTORY

Caitríona Murray



A thesis submitted to the University of Dublin

for the degree of Doctor of Philosophy

TRINITY COLLEGE
0 6 NOV 2008
LIBRARY DUBLIN

Smurfit Institute of Genetics

Trinity College

University of Dublin

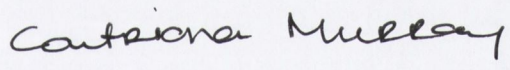
July 2008

Thesis
8641

Declaration

I hereby certify that this thesis has not been previously submitted for examination to this or any other university. The work described herein has been carried out by the author alone, except where otherwise stated.

This thesis may be made available for consultation within the university library and may be photocopied or loaned to other libraries for the purposes of consultation.



Caitriona Murray

July 2008

Summary

Domestication, and its associated environmental and demographic changes, undoubtedly brought new challenges to early domesticates, such as cattle. One of the most important would have been the emergence of new disease challenges and changes in disease epidemiology. If, as a result, selection acted on genes conferring increased disease resistance, for example, then signatures of this selection should be evident in the genomes of modern cattle. Cattle are economically important all over the world and losses due to disease are significant, particularly in the developing world. Identifying putative sites of selection helps to drive new discoveries by directing future work to the regions with most potential impact. This may be done "blind", using a genome-wide scan for unusual loci, or by targeting specific candidate genes based on prior knowledge. Here, both approaches are taken. By examining diversity at a particular locus in the context of wider genomic diversity, it should be possible to distinguish the signature of selection from the effects of demographic history.

Genome-wide variation in the parents of the USDA USMARC reference families was analysed using 3765 microsatellite in 28 individuals, with an average marker interval of 1.4cM. A Ewens-Watterson type approach was used to compare the observed heterozygosity to the expected heterozygosity under a model of neutral evolution. Departures from expected heterozygosity were observed, all with less heterozygosity than expected under neutrality. Results were consistent under different models of microsatellite mutation. A number of genes were identified flanking the outlying loci, which could be considered candidates for selection. However, given the small and imperfect population available for this study, these results must be interpreted cautiously.

Microsatellites were also used to assess the genetic variation around the KIR gene cluster, a candidate region for balancing selection. The close functional relationship between the KIR genes and their MHC ligands, coupled with analyses from the human KIR gene cluster, suggest that these genes are good candidates for selection. Seventeen microsatellites, located across an approximately 30cM region spanning the KIR region, were genotyped in 6 cattle populations representing cattle breeds from around the world. Patterns of genetic distance and deviations from expected heterozygosity were suggestive of balancing selection at these loci.

A resequencing approach was taken to investigate whether the genetic variation at candidate genes involved in immune response was consistent with non-neutral evolution. Candidate genes were sequenced in a panel of cattle representing European *Bos taurus*, African *Bos taurus* and *Bos indicus* and outgroup species. Haplotypes were reconstructed and deviations from neutrality assessed using a panel of standard neutrality tests. A coalescent approach was used to assess the significance of these deviations under a neutral model and under a model including a population expansion. The most robust indications of selection came from the two interleukin genes included in the study, IL2 and IL13. Exon 3 of CD2 was also highlighted as a possible site of selection. The majority of genes showed no significant deviations from neutrality.

The same sequence data were used to investigate the demographic history of cattle. Simulated populations were generated under a number of popular models of demographic history and Tajima's D for each simulation was compared to the observed data for goodness of fit. A coalescent-based Bayesian MCMC analysis of a subset of the sequence data were used to generate Bayesian Skyline Plots, which describe the effective population size of the sample backwards through time. Neither of these analyses support the hypothesis of a severe bottleneck at the time of domestication. Furthermore, the estimates of effective population size are considerably larger than previous estimates. Differences in the demographic history of *Bos taurus* and *Bos indicus* populations were also evident; the *Bos taurus* populations show evidence of a significant expansion in population size, while the *Bos indicus* data is more consistent with a relatively constant population size.

Acknowledgements

I would like to thank Dan for his patience, advice and encouragement, especially when the going got tough. I feel fortunate to have him as my supervisor.

To the members of the lab: Valeria, Ceiridwen, Emma, Brian, Lilian, Yonas, Frauke, Simone, Tim and Sarah as well as those who have moved on to greener pastures: Laoise, Ruth, Dave, Helge and Fernando. Their support and willingness to help has been invaluable. I'm not sure how I'll cope without the constant supply of sweet treats! Thanks for making the Bovine lab a great place to be. A special word of thanks to Ruth who gave me so much of her time over the past few years.

Thanks also to Brenda, Sue, Dave and Paul who keep things running smoothly in the department.

Genotype data used in Chapter 2 was kindly provided by Gary L. Bennett (USMARC) and Yoshikazu Sugimoto (Shirakawa Institute of Animal Genetics). Additional genotype data in Chapter 3 was provided by Stephen Park and Ruth Freeman. Ruth Freeman also provided additional sequence data used in Chapter 4 and Chapter 5.

My friends, both in the Genetics Department and outside in the real world, have been a constant source of encouragement and advice.

I am grateful for the support I received from my family. My parents have never stopped believing in me. They were there for me through thick and thin, and have a miraculous ability to know when not to ask.

Finally, I would like to dedicate this thesis to Rich. He has been there for me all the time, even when he wasn't *here*. From python lessons and proofreading to cups of tea and tissues, without his constant support and encouragement I would not have got this far.

Contents

Declaration i
Summary ii
Acknowledgements iv
Contents v
Figures viii
Tables ix

1 General Introduction 1
1.1 Cattle and Domestication 2
1.1.1 Domestication 2
1.1.2 Modern Cattle Breeds 5
1.1.3 Why look for selection in immune genes of cattle? 7
1.2 Evolution and Selection 9
1.2.1 The neutral theory of evolution 9
1.2.2 Selection 9
1.2.3 Demographic history 10
1.2.4 Searching for selection: genome scans and candidate gene approaches 11

2 Analysis of Genome-wide Microsatellite Diversity in an Incidental Population 13
2.1 Introduction 14
2.1.1 Microsatellites: mutation and evolution 14
2.1.2 Models of mutation and microsatellites 15
2.1.3 Ewens Watterson based approach 16
2.2 Materials and Methods 17
2.2.1 The dataset 17
2.2.2 Phylogenetic tree/MDS 17
2.2.3 Ewens Watterson based analyses of heterozygosity 18
2.2.4 Correlations between lab data and genome wide microsatellites. 18
2.3 Results 20
2.3.1 Dendogram 20
2.3.2 Analysis of microsatellite heterozygosity 21
2.3.3 Genes in the region of outlying markers 25
2.3.4 Correlation between Lab data and USDA USMARC data 34
2.4 Discussion 36
2.4.1 Sample population 36
2.4.2 Observed and Expected heterozygosity 36
2.4.3 Genes near the negative end of the distribution of DH/sd 37
2.4.4 Marker redundancy 38
2.5 Conclusions 39

3 Microsatellite Variation Around a Candidate Region for Selection 40
3.1 Introduction 41
3.1.1 Natural Killer cell receptors and the immune response 41
3.2 Materials and Methods 47
3.2.1 DNA samples 47
3.2.2 Microsatellites 47

3.2.3 PCR amplification and genotyping 49
3.2.4 Data analyses 52
3.2.5 Additional data 55
3.3 Results 56
3.3.1 Data generated 56
3.3.2 Genomic positions of the microsatellites. 57
3.3.3 HWE 59
3.3.4 Gene Diversity 61
3.3.5 Genetic Distances 61
3.3.6 Ewens Watterson based tests (H_{OBS}/H_{EXP} and DH/sd) 66
3.4 Discussion 68
3.4.1 Comparison of BTA18 microsatellite data with similar data surrounding the MHC. 68
3.4.2 Comparison with a genome-wide microsatellite dataset 71
3.5 Conclusions 75

4 Investigation of Candidate Genes by Sequence Analysis 77
4.1 Introduction 78
4.1.1 The Genes 78
4.1.2 The Coalescent 84
4.2 Materials and Methods 86
4.2.1 Screening sets 86
4.2.2 Primer design for resequencing 89
4.2.3 PCR and Sequencing 89
4.2.4 Additional Data 94
4.2.5 Sequence Analysis Pipeline 95
4.2.6 Networks 97
4.2.7 DnaSP 97
4.2.8 Haplotype Diversity 97
4.2.9 Neutrality Tests 98
4.2.10 Assessing significance by coalescent simulations 101
4.3 Results 104
4.3.1 Sequencing 104
4.3.2 Polymorphic sites 107
4.3.3 Haplotypes 122
4.3.4 Linkage Diseqilibrium 135
4.3.5 Neutrality Tests 146
4.3.6 Sliding window Analyses 159
4.3.7 Assessing significance of neutrality tests with coalescent simulations 175
4.4 Discussion 182
4.4.1 Interleukins 183
4.4.2 CD2 185
4.4.3 TRIF 186
4.4.4 PRF1 and Toll-like Receptors (TLRs) 186
4.5 Conclusions 190

5 Coalescent Simulation and Bayesian Analyses of Demographic Histories 193
5.1 Introduction 194
5.1.1 Neutrality tests and demographic history 194
5.1.2 Bayesian Skyline Plots 194

5.1.3 Analyses of bovid DNA using skyline plots 195
5.2 Materials and Methods 197
5.2.1 Coalescent simulations 197
5.2.2 BEAST 199
5.3 Results 202
5.3.1 Coalescent-based analysis (MS) 202
5.3.2 Bayesian analysis (BEAST) 207
5.4 Discussion 213
5.4.1 Different demographic histories for continental populations 213

6 Conclusions 219

7 References 222

Appendix A 232

Figures

Figure 2.1 Neighbor joining tree constructed from a distance matrix based on allele sharing using the Neighbor application of Phylip (Felsenstein 1989). 20
Figure 2.2 Multidimenstional scaling (MDS) showing the distance between samples constructed from the allele sharing distance matrix. 21
Figure 2.3 Departures from expected heterozygosity. 23
Figure 2.4 Genome wide distribution of DH/sd 24
Figure 2.5 Correlation between DH/sd of 77 markers genotyped in three populations.... 35
Figure 3.1 A schematic diagram illustrating the expansion of the KIR genes located in the LRC on human chromosome 19, and of the Ly49 genes located in the NKC on mouse chromosome 6. (Taken from Trowsdale et al. 2001) 42
Figure 3.2 Phylogenetic tree constructed using amino acid sequences of individual Ig domains from bovine (bt) and selected primate (h) KIR genes. 45
Figure 3.3 Schematic of microsatellites successfully amplified from BTA18q. 56
Figure 3.4 Gene diversity for 14 loci on BTA18. 60
Figure 3.5 Genetics distances between populations at 14 loci on BTA18. 62
Figure 3.6 Allele frequency distributions for DIK5325 (left) and BM2078 (right). 64
Figure 3.7 Ratio between the observed and expected heterozygosity. 65
Figure 3.8. DH/sd plotted against chromosomal position. 67
Figure 3.9 F_{OBS}/F_{EXP}, genetic distance and gene diversity at 18 loci on a region BTA23 encompassing the MHC 70
Figure 3.10 Distribution of DH/sd for each of the six breeds included in the study. 72
Figure 3.11 DH/sd distributions across all populations subdivided into non-KIR/MHC (top), KIR-only (middle) and MHC-only (bottom) 73
Figure 4.1 Overview of the sequence analysis pipeline used in this study. 96
Figure 4.2 Models of demographic history 102
Figure 4.3 Schematic diagram illustrating the structure of each of the genes included in this study and the regions sequenced 104
Figure 4.4 Alignments of polymorphic sites. 107
Figure 4.5 Reduced median (RM) networks. 123
Figure 4.6 Significant pairwise linkage disequilibrium (LD). 135
Figure 4.7 Variation in π, divergence between continental populations (D_a) and Tajima's D across each gene. 159
Figure 5.1 Models of demographic history simulated using MS (Hudson, 2002). 198
Figure 5.2 Distribution of (Dobs-Dsim)/SDsim for each of the four models of demographic history. 204
Figure 5.3 Bayesian Skyline Plots for (a) *Bos indicus*; (b) All *Bos taurus*; (c) African *Bos taurus*; (d) European *Bos taurus*; (e) Reduced European *Bos taurus* data (n=11). .212

Tables

Table 2.1 Breed, sex and relatedness information for all animals in the sample population. 17
Table 2.2 Microsatellite loci at the negative tail of the DH/sd distribution. 25
Table 2.3 Map positions of loci at the negative end of the distribution. 25
Table 2.4 Genes within ± 0.5Mb of markers at the negative end of the distribution of DH/sd. 27
Table 3.1 PCR components and concentrations for amplification of microsatellite loci.. 50
Table 3.2 PCR conditions, primer pairs and map information for primer pairs included in this study. 51
Table 3.3 Microsatellite loci used in study. 57
Table 3.4 Genomic positions of microsatellites used in study 58
Table 3.5 Deviations from Hardy-Weinberg Equilibrium 59
Table 3.6 Mann-Whitney U-Test of different DH/sd distributions in the six breeds 74
Table 4.1 Breed information for panel A 87
Table 4.2 Breed information for panel B. 88
Table 4.3 Primer sequences and concentrations, $MgCl_2$ concentrations, annealing temperatures. 90
Table 4.4 Haplotypes and haplotype diversity for each gene in each sample populations (3d.p.). Table continues on following page 133
Table 4.5 Nucleotide diversity and Tajima's *D* values obtained for each gene in each sample population. Table continues on following page. 147
Table 4.6 Fu and Li's *D* and *F* (tests with an outgroup) values obtained for each gene in each sample population. Table continues on following page. 150
Table 4.7 Fu and Li's *D** and *F** (tests without outgroups) values obtained for each gene in each sample population. Table continues on following page 152
Table 4.8 Fay and Wu's *H* values obtained for each gene in each sample population. Table continues on following page 155
Table 4.9 Fu's F_S, observed value and probability for each gene in each sample population. Table continues on following page. 157
Table 4.10 Probability of observed Tajima's *D* and Fay and Wu's *H* and confidence intervals determined by coalescence analysis using MS. Table continues on following page 177
Table 4.11 Windows identified as having significant values of Tajima's *D* (D), Fu and Li's *F* (F*)* and Fu and Li's *D* (D*)* by coalescent simulation under a) SNM and b) EXP1 model. Table continues on following pages 180
Table 4.12 Summary of neutrality test results for IL2 and IL13. 191
Table 5.1 Summary of (Dobs-Dsim)/SDsim for each of the four models for each population. 205
Table 5.2 Wilcoxon Signed Rank test results. 207
Table 5.3 Modeltest results 208

1 General Introduction

1.1 Cattle and Domestication

Cattle are an economically important domesticated animal, perhaps the most economically important with the global herd estimated to be in excess of 1.3 billion (Bradley and Cunningham 1999). They are typically used for dairy, beef and/or draft purposes. They are economically important and in many cultures are a symbol of wealth and social standing. Modern cattle are classified into two morphological categories often called 'zebu' and 'taurine'.

Zebu, the cattle typical of the Indian subcontinent, have a characteristic hump and dewlap that is not found in taurine cattle. They are more suited to arid environments, having larger sweat glands and greater productivity under sub optimal conditions. Furthermore, they have a greater resistance to a number of diseases, particularly rinderpest, ticks and tick-borne diseases, and gastrointestinal parasites (Bradley and Magee 2006).

Bos taurus, cattle of European or African origin show the taurine morphology, while the *Bos indicus* of Asia show the zebu morphology. Migrations and introgression of *Bos indicus* cattle into *Bos taurus* populations has resulted in a blurring of the dichotomy of morphology and it is not uncommon to find genetic signatures typical of *Bos indicus* in morphologically taurine animals, particularly in the Near East and Africa, where introgression is common.

1.1.1 Domestication

Modern cattle were domesticated from the wild ancestor *Bos primigenius*, commonly known as aurochs. Aurochs ranged across Europe, Northern Africa and much of Asia, except Western China and the Himalayas and the colder north of the Eurasian landmass. Regional subspecies existed, *Bos primigenius primigenius,* the European aurochs, *Bos primigenius opisthonomus*, the African aurochs and *Bos primigenius namadicus*, the Asian aurochs. The last aurochs, a European aurochs, died in Poland in the 17th century AD. Domesticated cattle and their wild progenitors must therefore have coexisted for millennia.

The earliest evidence for domestication of cattle is found in the region known as the Fertile Crescent, in the Near East. The Fertile Crescent ranges from the Persian Gulf in the East to the Mediterranean Sea in the west and is a putative domestication centre for numerous animal and crop species. Prior to the advent of population genetic analyses, it was widely thought that a single domestication event in the Near East was a common origin of all domesticated cattle (Epstein 1971; Epstein and Mason 1984) and from there

expanded eastward into Asia, north into Europe and westward into Africa. It was thought that the morphological differences between *Bos indicus* and *Bos taurus* were the result of subsequent selection during breeding. Genetic analyses have fundamentally disproved this hypothesis. Alternative hypotheses included a local domestication of *Bos indicus* from *Bos primigenius namadicus* progenitors in South Asia, as evidenced by archaeological findings in Baluchistan (Meadow 1993). More controversial are the suggestions of a local domestication of African *Bos taurus* from *Bos primigenius opisthonomus* in Northern Africa between 8000 and 9000 years ago (Wendorf and Schild 1994) and local domestications or introgression of European *Bos primigenius primigenius* into European *Bos taurus* populations (Bökönyi 1974). It should be noted that these hypotheses are not mutually exclusive. As advances in ancient DNA techniques allow more data to be retrieved from archaeological samples, these hypotheses can be further explored.

Analysis of mitochondrial DNA from European, African and Asian cattle revealed a pronounced dichotomy in the phylogenetic tree of extant cattle. African and European *Bos taurus* formed one clade, while *Bos indicus* formed the other highly divergent clade. The time since the divergence of these two clades was estimated to be in the order of 200,000 years, long before the domestication of cattle occurred. This is an indication that at least two domestication events occurred, with *Bos taurus* being domesticated from *Bos primigenius primigenius* and *Bos indicus* being domesticated from *Bos primigenius namadicus*.

A third centre of domestication has been suggested in Northern Africa, with *Bos primigenius opisthonomus* being the progenitor species. Mitochondrial and microsatellite data support this hypothesis (Bradley et al. 1996; Hanotte et al. 2002). Estimates from mitochondrial DNA data put the time of divergence between African cattle and European cattle in the region of 22,000 - 26,000 years ago (Bradley et al. 1996), which was before the earliest domestication approximately 10,000 years ago. Microsatellite data yield a much older estimated divergence of 150,000 to 250,000 years ago (MacHugh et al. 1997). Troy et al. (2001) analysed mitochondrial DNA haplotypes in European, Near Eastern and North African populations. In addition to observing star-like phylogenies, they found that the most common haplotype in Northern African populations was present in the Near East - the putative centre of domestication of all *Bos taurus* - at only low frequencies. Furthermore, this haplotype was completely absent from most European populations

though it had been reported elsewhere to be found in Portuguese cattle (Cymbron et al. 1999).

The hypothesis of an African domestication is still controversial. It has been argued that a migration of a small number of animals out of the Near East into Northern Africa is a more parsimonious explanation, and that the high frequency of the common mitochondrial haplotype in African cattle can be explained by a simple founder effect (Beja-Pereira et al. 2006).

Genetic diversity in African cattle has been deeply influenced by migrations of *Bos indicus* into Africa. The earliest cattle in Africa were *Bos taurus*, but today there are numerous breeds with zebu morphology. Many breeds, including morphologically taurine breeds, show genetic evidence of *Bos indicus* ancestry (Hanotte et al. 2002; MacHugh et al. 1997).

Emerging genetic data suggest that the process of cattle domestication has been more complex than the relatively simple domestication, expansion and migration model that is commonly considered. Several lines of evidence suggest that introgression or back crossing with the progenitor species, aurochs, is likely to have occurred (Beja-Pereira et al. 2006; Götherström et al. 2005; Vila et al. 2005).

Analysis of mitochondrial DNA from aurochs teeth recovered in Italy places these Italian aurochs within the clade representing modern European cattle (Beja-Pereira et al. 2006). This is in contrast to earlier analysis of British aurochs samples, which suggested that those animals formed a separate clade, divergent from modern *Bos taurus* and as such, likely made no contribution to the modern *Bos taurus* population (Troy et al. 2001). This was taken as evidence that modern European cattle were domesticated in the Near East. Overall, modern cattle have very low mitochondrial diversity relative to other ungulates (Bradley and Magee 2006). This pattern of reduced diversity in the mitochondrial genome is not repeated in the nuclear genome. High diversity was reported at MHC loci in a number of domesticated species (Vila et al. 2005). Balancing selection at the MHC results in alleles being maintained in the population for long periods of time, but given the relatively short time since domestication, it was determined that the vast majority of alleles observed at the MHC in domesticated species must have existed in the ancestral population. Vila et al. suggest that backcrossing with the wild progenitor species offers a plausible explanation for the observed high diversity at nuclear loci.

Y chromosome data from European cattle also supports this theory (Götherström et al. 2005). The dominant Y chromosome haplotype in northern European cattle was found to be the major Y chromosome haplotype identified in European aurochs or early domestic cattle with 20 out of 21 samples having this haplotype. This haplotype was found to be much less frequent in southern European cattle. Domesticated cattle and their wild progenitor, aurochs, coexisted for thousands of years. Furthermore, there is little archaeological evidence of fencing suggesting that early domesticated cattle were relatively free roaming. Such a scenario would offer ample opportunity for hybridisation with wild aurochs. Mating between a domesticated cow and wild male would introduce 'wild' nuclear DNA and Y chromosome into the domesticated population, as the offspring would most likely be retained in the domesticated herd. In the opposite scenario, a mating between a wild female and a domesticated male, the offspring would be less likely to be incorporated into the herd, meaning that 'wild' mtDNA would be less likely to be incorporated into the gene pool of domesticated cattle. This would explain the observed results from Y chromosome analyses as well as the apparent absence of aurochs mitochondrial DNA from the heavily sampled European *Bos taurus* population.

The theory of introgression from wild aurochs remains controversial. Edwards *et al.* (2007) found no indication of European aurochs ancestry in modern populations of cattle from northern and central Europe. They postulate that the results of Beja-Pereira *et al.* (2006) could be due to the Italian aurochs samples being part of a glacial refugia population that never crossed the Alps after the end of the last ice age.

1.1.2 Modern Cattle Breeds

Cattle breeds as they are known today have been formally recognised and bred for at most a few hundred years and many are more recent. There are an estimated 800 cattle breeds (Bradley and Cunningham 1999), with *Cattle Breeds: An encyclopedia (Felius 1995)* offering a description of almost 700 breeds. DNA samples from a range of breeds have been used in the work presented here and are described below. These breeds were chosen to represent cattle populations of African *Bos taurus*, European *Bos taurus* and *Bos indicus*.

African breeds with low levels of *Bos indicus* introgression have been selected to represent African *Bos taurus*. These breeds are N'Dama, Somba and Lagune. N'Dama are a longhorn breed found all across western Africa as a result of their resistance to trypanosomiasis, or sleeping sickness, a disease transmitted by the tsetse fly, which is endemic in Western Africa. The most pure N'Dama are Guinean N'Dama,

which have been shown to have extremely low levels of *Bos indicus* admixture (0.02%) (MacHugh et al. 1997). Somba are small West African shorthorn breed that originate in the Atacora highlands of northern Benin. The samples used here were sampled in Benin and neighbouring Togo. Lagune are also a West African shorthorn breed, but are likely derived from the larger Savannah shorthorn. This breed is found in coastal areas of the Gulf of Guinea *(Felius 1995).*

European *Bos taurus* are presented by smaller samples from a larger number of breeds. Friesian cattle have a characteristic black and white pied coat and originated in Friesland in north of Holland. They are large dairy cattle and are economically important. They have been exported worldwide and are favoured because of very high milk yields. Closely related is the German Black, which originated in northwestern Germany in the region neighbouring Friesland. Charolais originated in central France, and are likely descended from same ancestors as Simmental. Charolais is a globally important beef breed. Simmental are indigenous to the Simme and Saane valleys of the canton of Bern in Switzerland. They are bred for dairy and beef. Gelbvieh are closely related to Simmental, originated in Central Germany. They are an important beef breed, especially in the USA. Aberdeen Angus is an important beef breed whose ancestry traces back to polled cattle in Aberdeenshire in 19th century Scotland. The other Scottish breed in the study is Highland. This is a very old breed from the highlands of Scotland, easily identified by its shaggy coat. Hereford originated in Herefordshire in England. It is known to have some ancestry with Dutch or Flemish cows. Norwegian Red is a recently founded breed and has been described as a heterogeneous population of cattle. It has been bred with both native and imported animals including Friesian and Ayrshire. Sikias is a Greek Steppe breed that is critically endangered with only isolated populations remaining on Khalkidiki, the Sithonia peninsula and the island of Peperi. Romagnola originated in Emilia in Italy and is one of a number of white breeds in Italy. Two Portuguese breeds are included. Alentejana, from Alentejo in southern Portugal and Mertolenga, a relatively recently formed breed which has Alentejana ancestry. There have been shorthorn cattle in Anatolia for more than 4000 years *(Felius 1995).*

Bos indicus are represented by Sahiwal, Hariana and Tharparker from the Northern part of the subcontinent and Ongole and Nellore from southern India. Sahiwal have been bred in the Punjab region of northern Pakistan for many years. It has been suggested that the ancestral cattle originated in Northern India, northern Pakistan and Afghanistan. They have very good resistance to heat and have been exported all over the world. Hariana

originate in the fertile plains of Haryana state in northern India. They are well adapted to semi-arid sub-tropical regions. Tharparker are named after the Thar Desert in western India and southeastern Pakistan. The breed is thought to have originated in this region. Ongole is another shorthorn Indian breed, originating in the Guntur-Nellore region of Andhra Pradesh in the south of the country. The Nelore is a Brazilian breed derived from the Ongole and taking its name from the region where the Ongole originated *(Felius 1995)*.

These breeds are representative of global cattle diversity, with the exclusion of known *Bos indicus-Bos taurus* hybrids such as the African 'sanga' breeds.

1.1.3 Why look for selection in immune genes of cattle?

The global cattle population exceeds 1.3 billion making the cattle industry one of the most economically important industries across much of the world. Losses due to disease are significant and the study of disease resistance has long been of interest. Identifying loci involved in disease resistance could potentially help improve the treatment and management of cattle pathogens. Different cattle populations have encountered different disease challenges and exhibit different innate resistances. West African breeds such as N'Dama are more resistant to trypanosomiasis while *Bos indicus* breeds are more resistant to a range of parasites including ticks and gastrointestinal parasites. Direct evidence that a locus has undergone positive selection in recent bovine history is one way to identify those loci that are of most significance in resisting diseases encountered during and after domestication

Evidence from phylogenetic analysis of human pathogens indicates that a number of human diseases (i.e. measles and pertussis) may have arisen as a result of domestication (Pearce-Duvet 2006). Interestingly, such analysis has also suggested that domestication was not the source of a number of other human diseases often attributed to domestication. The bacterium that causes human tuberculosis (TB), *Mycobacterium tuberculosis* is highly divergent from the bacterium that causes bovine tuberculosis, *Mycobacterium bovis*, and this divergence long predates domestication.

Domestication resulted in significant changes to both the population structure and environment of cattle. Higher population density and closer proximity to both humans and other domesticated species must have altered the epidemiology of infectious diseases in cattle. Cross-species transmission as a result of contact with other species would have introduced novel diseases into the population. Migration and environmental changes

brought about by early farmers would also have exposed cattle to new disease challenges. Stable isotope data from cattle and aurochs bones suggests that aurochs and cattle had different feeding patterns, likely reflecting a movement from foraging in forests to grazing in open grasslands (Noe-Nygaard et al. 2005). Higher population densities would provide a larger population of hosts for pathogens.

Cattle have undergone numerous changes during domestication and a legacy of those changes will have been left in the genetic diversity of modern domestic populations. It is therefore possible that the large shifts in selective pressures experienced during domestication and exposure to new environments and new disease challenges will have left a detectable signature of selection on genes which have been important in disease resistance. Identifying these loci is an important step in combating the disease challenges faced by modern cattle, particularly in the developing world. Selectively breeding disease resistance into these populations has obvious benefits over veterinary medicinal approaches. Identifying resistance-enhancing loci provides a foundation for such breeding programs.

1.2 Evolution and Selection

1.2.1 The neutral theory of evolution

In the late 1960s the neutral theory of evolution was proposed by Motoo Kimura (Kimura 1968). The Neutral theory challenged the selectionist theories of the time, which held that positive and balancing selection were the primary forces controlling genetic variation and evolution. The Neutral theory proposed that the majority of new mutations are selectively neutral or nearly neutral, that is, they have a negligible effect on the fitness of the carrier. Therefore, random genetic drift is the major force governing the fate of such mutations. Under the neutral theory the most likely fate of a new mutation is extinction, though a small proportion of mutations will become fixed in the population purely as a result of drift.

However, the neutral theory does not preclude the action of selective forces on the genome. Clearly not all mutations that arise in a population are selectively neutral. A minority of mutations are either advantageous or deleterious, resulting in increased or decreased fitness of an individual carrying them, under particular environmental conditions. In the case of such mutations, natural selection drives these non-neutral mutations to fixation or extinction more rapidly than would be expected under neutrality.

1.2.2 Selection

Prior to the proposition of the neutral theory, balancing selection was thought to be the major force maintaining polymorphism in the genome. It is now considered likely that it plays a much more limited role (Asthana et al. 2005; Kreitman and Di Rienzo 2004). Balancing selection acts to maintain multiple polymorphisms at a given locus in the population and consequently can be identified by an excess of intermediate frequency alleles compared to expectation under neutrality. As balancing selection favours the retention of polymorphism at a particular site, it can result in elevated diversity within populations and reduced diversity between populations. Even very low levels of migration between populations can result in reduced diversity between populations, as migration will introduce new alleles, which will be selectively maintained in the population. Furthermore, as a result of balancing selection, alleles can persist at intermediate frequency in the population for long periods of time, and can even be maintained in two species after speciation. The most well documented case of balancing selection is that of the Major Histocompatibility Complex (MHC). The MHC genes, refered to as Human Leukocyte Antigens (HLA) are among the most polymorphic genes

in the human genome, with over 560 HLA-B alleles described to date (Luo et al. 2005). Furthermore, at least 11 trans-species polymorphisms at MHC loci are shared between human and chimpanzee, whose lineages diverged in the region of 5-7 MYA (Asthana et al. 2005).

Positive selection of a mutation conferring a selective advantage will increase the frequency of the allele in the population. In addition, linked neutral loci are also affected in a process called hitchhiking, resulting in reduced levels of diversity at and around the locus under selection. The strength of the selective force depends on how beneficial the mutation is to the individual under the current environmental conditions. For example, resistance to pyrimethamine/sulfadoxine (PS) as an anti-malarial treatment on the Thailand-Myanmar border spread to fixation in 36 generations of the malarial parasite *Plasmodium falciparum*, which equates to approximately six years, following the introduction of PS as the primary treatment of malaria in the mid 1970s. Genotyping of 5 microsatellites flanking the dihydrofolate reductase gene (*dhfr*), the locus known to be responsible for resistance to pyrimethamine in *Plasmodium falciparum*, revealed an extensive region of reduced diversity covering approximately 6cM around the *dhfr* gene (Nair et al. 2003).

It is assumed that the majority of non-neutral mutations that arise in a population are deleterious, rather than advantageous. Negative selection, also referred to as background or purifying selection acts to eliminate these deleterious alleles from a population, and is predicted to have a similar, though less pronounced, effect on diversity as positive selection. Negative selection is also likely to be a more common event than either positive selection or balancing selection given that the majority of non-neutral mutations arising in the genome are likely to be deleterious.

1.2.3 Demographic history

The ability to detect the effect of selection by examining genetic diversity is hindered by the fact that changes in the demography of a population can have a similar effect on genetic diversity.

Natural populations are rarely constant. A population bottleneck, where the population size is severely reduced for a period of time, can dramatically reduce the genetic diversity of the population. The signatures of a recent, severe bottleneck, from which the population has recovered, including reduced levels of diversity and an excess of low frequency alleles, are the same as what would be expected from positive selection.

Similarly an expanding population would also exhibit an excess of low frequency alleles. Population structure, which might arise if a population is formed from two distinct subpopulations, or as a result of even moderate levels of migration between populations over long periods of time, could result in an excess of intermediate frequency alleles, much like that expected as a result of balancing selection.

Although selection and changes in population demography can have the same effect on genetic diversity there is a key difference that enables the study of genetic diversity to identify regions of the genome that have been subject to selection. While the effect of selection is limited to the site and flanking regions of the genome the effect of demographic history is spread equally across the whole genome. Thus by examining diversity at a particular locus in the context of wider genomic diversity, it should be possible to distinguish the signature of selection from the effects of demographic history.

1.2.4 Searching for selection: genome scans and candidate gene approaches

Identifying genes that have been subject to the effects of selection is of great interest for a number of reasons. Identifying such genes could contribute to a number of lines of research into gene function, disease mechanisms, or disease resistance. Discovering our past, and the past of those species that live closely with us, including domesticates such as cattle, is inherently interesting. By identifying genes that have been selected for in the evolution of these species, we can hope to infer something of the selective pressures that our ancestors faced and in doing so we can learn more about where we came from. Genes involved in the immune response have been found to be over-represented in a number of studies that aimed to detect evidence of selection (Nielsen et al. 2007). This makes sense given that genes targeted by selection must have some functionally significant effect on the fitness of the individual in the population.

A range of approaches have been used to attempt to detect signatures of selection. In addition to inter-species analyses (Clark et al. 2003; Nielsen et al. 2005) much work has been carried out at the population level. Genome scans making use of the increasing availability of genetic data in the form of SNP data (Akey et al. 2002; Carlson et al. 2005; Voight et al. 2006), large scale resequencing data (Akey et al. 2004; Bustamante et al. 2005) and microsatellite data (Payseur et al. 2002). From studies like these, it is becoming clear that robust identification of selected loci requires large numbers of loci and samples sizes. This is an exciting time to look for signs of selection as improvements

in technology are making the availability of these data a reality. In addition, new methodologies improve our ability to model the data and extract the signal from confounding effects of demographic history.

2 Analysis of Genome-wide Microsatellite Diversity in an Incidental Population

2.1 Introduction

2.1.1 Microsatellites: mutation and evolution

The selection of traits of production or behavioural importance has been a central theme in cattle history. This has been increasingly stringent in recent centuries but is likely to be a process stretching back to the earliest husbandry. Additionally, genes involved in the immune response are also likely to have been under selective pressure following domestication and the radical change in demography associated with it; increased proximity to animals of the same and other species, combined with habitat changes would have brought novel disease challenges. Population genetic signatures of past selection may be inferred from regions of the genome that show outlying patterns of genetic diversity and an exciting and timely project is to search for such regions within the cattle genome, perhaps as an aid to identification of genes of significance (Luikart et al. 2003).

Microsatellites consist of tandemly repeated units, typically one to six nucleotides in length. They are highly polymorphic, abundant in mammalian genomes and are valuable tools for linkage mapping and population genetics. They are assumed to evolve neutrally. Given this, identifying microsatellites whose pattern of variation deviates from the expectation may be a useful and cost effective method of identifying regions of the genome that have been influenced by selection, without the need for prior knowledge of candidate genes (Kayser et al. 2003; Payseur et al. 2002).

Positive selection acting on an advantageous allele increases the frequency of that variant in the population. In addition, linked neutral loci are also affected in a process known as hitchhiking, resulting in reduced levels of diversity at and around the selected site. Balancing selection, in contrast, will result in elevated diversity levels in the region affected.

The ability to detect signatures of selection is confounded by demographic history, which may have similar effects on genetic variation as that of selective forces. However, while the effect of selection is limited to the specific gene and flanking regions, the effects from demographic history influence patterns of diversity across the whole genome. Thus by assessing loci in the context of genome-wide pattern of variation, it should be possible to distinguish the signature of selection from these more general effects. An empirical approach also restricts reliance on analytically derived significance levels, which may be misleading (Schlotterer et al. 2004).

2.1.2 Models of mutation and microsatellites

In order to order to analyse microsatellite diversity for deviations from neutrality, it is important to have a model of microsatellite mutation with which to make predictions. Due to their repetitive nature, the primary mechanism of microsatellite mutation is probably slippage during DNA replication, resulting in the gain or loss of nucleotide repeats. There are a number of different models for microsatellite evolution. (see Ellegren 2004) for a review of microsatellite evolution). The software used for this study, BOTTLENECK (Cornuet and Luikart 1996), implements three models that differ primarily in terms of the number of repeat units that are gained or lost with each mutational event. The Infinite Allele Model (IAM) (Kimura and Crow 1964) is the simplest. It does not explicitly model repeat numbers and results in the creation of a new allele after every mutational event. There are therefore no recurrent or back mutations. Under the Stepwise Mutation Model (SMM) (Ohta and Kimura 1973) the mutational events involve the gain or loss of a single repeat unit only. The Two Phase Model (TPM) (Di Rienzo et al. 1994) may be a more realistic model than either the IAM or the SMM. According to the TPM, the majority of mutational events involve the gain or loss of a single mutational unit, while a minority of events result in a larger change.

Recent work on cereal microsatellite diversity has suggested that neither the TPM nor the SMM are the most appropriate models for microsatellite evolution in these species (Casa et al. 2005; Matsuoka et al. 2002). From a sample of 72 microsatellites that were polymorphic in *Sorghum bicolour* (sorghum), only 16 loci (23%) had an observed allele size distribution consistent with either the TPM or the SMM (Casa et al. 2005). Earlier work on microsatellite diversity in a number of *Zea* species (including maize and its wild relative teosinte) suggested that this percentage might be even lower for maize species (Matsuoka et al. 2002). By directly sequencing a subset of alleles from six loci showing a range of allele size distribution patterns, Matsuoka and colleagues showed that insertion/deletions in the regions flanking the repeat unit are a major contributor to observed diversity at microsatellite loci. Despite these observations, the microsatellite data were informative for inter- and intra-species analysis. In the absence of any other model, the IAM appears to be more appropriate than either the SMM or TPM, though a model incorporating features of the IAM and SMM would likely be more appropriate. The extent to which indels occur in regions flanking mammalian microsatellites is unknown, but previous work on bovine microsatellites suggested that the most appropriate model should incorporate features of the IAM and SMM and that in the

absence of such a model, the IAM may be more useful (Park 2002). Because of these discrepancies, it is often useful to look at several different models to assess the degree to which observed features of the data are due to the choice of model.

2.1.3 Ewens Watterson based approach

The Ewens-Watterson test is based on Ewens' sampling theory, which shows that the sample size and the number of alleles observed within that sample are sufficient to determine the expected allele configuration under a model of neutral evolution (Ewens 1972). Based on this observation, Watterson developed a test of neutrality in which the expected allele configuration assuming neutrality is compared to the observed allele configuration in order to identify departures from neutrality (Watterson and Anderson 1978).

One way to display the results of this type of analysis in a large scale study is to plot the genome wide distribution of the values (e.g. DH/sd, see section 2.2.3). In such a distribution, the loci at the negative tail have lower observed heterozygosity than expected heterozygoisty and the loci at the positive end of the distribution have greater observed heterozygosity than expected. Lower observed heterozygoisty might result from positive selection, because this would result in the selected allele being at high frequency, the reduction of variability in the region and an excess of rare variants. Therefore, a lower number of heterozygotes than expected are observed. Conversely, when the observed heterozygosity is greater than the expected heterozygosity, this might be due to balancing selection maintaining multple alleles at high frequency resulting in few rare variants and an excess of heterozygotes.

We take such an approach here using marker data from the current bovine linkage map. This is based on a set of 3802 microsatellites genotyped in the USDA USMARC cattle reference families and provides high density coverage of the bovine genome with an average marker interval of 1.4cM (Ihara et al. 2004). By using genotypes from the mapping pedigree founder animals, we conduct a genome-wide survey of microsatellite variation. Our aim was to identify regions of the genome exhibiting deviations from expected patterns of diversity that may harbour potential candidates for selected genes that warrant further study.

2.2 Materials and Methods

2.2.1 The dataset

Genotype data were obtained for the parents of the USDA USMARC reference families so as to maximise the number of individuals in the sample, while minimising the relatedness between those individuals. This population is made up of 24 *Bos taurus* animals and four *Bos taurus* X *Bos indicus* animals (Table 2.1) and also includes six pairs of half sibs. Computationally expensive methods exist for estimating allele frequencies from samples with pedigree structure (McPeek et al. 2004). However, Broman (2001) calculates that the simple use of all data when both sibs are included in a sample gives an unbiased estimator of allele frequency, which is superior to that of including data from only one member of each sibship. Here I have chosen to retain both members of each of the 6 half-sib pairs.

Table 2.1 Breed, sex and relatedness information for all animals in the sample population.

Breed[1]	Male	Female	Half sib pairs[2]
Hereford	1	6	1
Piedmontese X Hereford	0	2	0.5*
Longhorn X Hereford	0	2	0.5**
Angus	0	7	2
Piedmontese X Angus	0	2	0.5*
Longhorn X Angus	0	2	0.5**
Gelbvieh X Simmental	2	0	1
Nellore X Hereford	0	2	0
Brahman X Angus	1	0	0
Brahman X Hereford	1	0	0
Total	5	23	6

1. Nellore and Brahman are *Bos indicus* breeds. The remaining breeds are all European *Bos taurus*.
2. A Piedmontese X Hereford and a Piedmontese X Angus share a parent(*), and likewise a Longhorn X Hereford and a Longhorn X Angus share a parent(**). For the remaining half-sib pairs, each of the pair is in the same breed group.

2.2.2 Phylogenetic tree/MDS

Pairwise genetic distances were calculated between all 28 animals based on allele sharing at autosomal microsatellites, such that the distance between any two individuals is given by (1-*Ps*) where *Ps* is the number of shared alleles summed over loci divided by twice the number of loci compared (Bowcock et al. 1994).

A neighbour joining tree was constructed from this distance matrix using the Neighbor application of the Phylip package (Felsenstein 1989).

Multi Dimensional Scaling (MDS) analysis of the same allele sharing matrix was carried out using SPSS (version 11, SPSS Inc.).

2.2.3 Ewens Watterson based analyses of heterozygosity

The approach taken here to assess allelic diversity patterns derives from the Ewens-Watterson test. In the present study heterozygosity, rather than homozygosity, has been used as a summary statistic to describe the allele configurations, making this approach effectively a reverse Ewens-Watterson test (Payseur et al. 2002).

This approach, implemented in BOTTLENECK (Cornuet and Luikart 1996) was used to estimate the expected allele frequency configuration and thus expected heterozygosity (H_{eq}) of each locus under the assumption of neutrality. For each locus in each population, the average H_{eq} was estimated from a distribution of expected heterozygosity, which was generated by 1000 replicate simulations, given the sample size and number of alleles at that locus, and assuming alternately, the infinite alleles model of mutation (IAM), the stepwise mutation model (SMM) and an intermediate model of microsatellite evolution, the two phase model (TPM). From this distribution the standard deviation (sd) of H_{eq} was also obtained. This was used to scale the difference (DH) between the observed heterozygosity (calculated directly from the genotype data) and the expected heterozygosity (estimated from the simulations). The resulting standardised measure of deviation from expected genetic diversity (DH/sd) is negative when the observed heterozygosity is lower than the expected heterozygosity, as would be expected following a selective sweep near the locus. DH/sd is positive where the observed heterozygosity is higher than expected, as might result from balancing selection. Given that model-based simulations may be not be comparable across loci of differing allele number (Schlotterer et al. 2004) we also took a fourth, empirical approach to determining DH/sd. For each locus with n alleles, DH/sd was calculated as the difference between the observed heterozygosity at that locus and the mean observed heterozygosity for all loci with n alleles divided by the standard deviation for all loci with n alleles.

2.2.4 Correlations between lab data and genome wide microsatellites.

Seventy seven of the microsatellites for which genotype data were obtained from the USMARC cattle reference families had previously been genotyped in a sample population of Charolais and Friesian, (Park, 2002 and references therein). DH/sd was calculated for each marker in each of these populations and the values obtained in the

three different populations were compared. The Pearson correlation and significance was calculated using SPSS (version 11, SPSS Inc.).

2.3 Results

Thirty-seven microsatellites with a high proportion of missing genotypes were excluded from the study leaving a set of 3765 microsatellites, 93 of which were monomorphic within the *Bos taurus* animals.

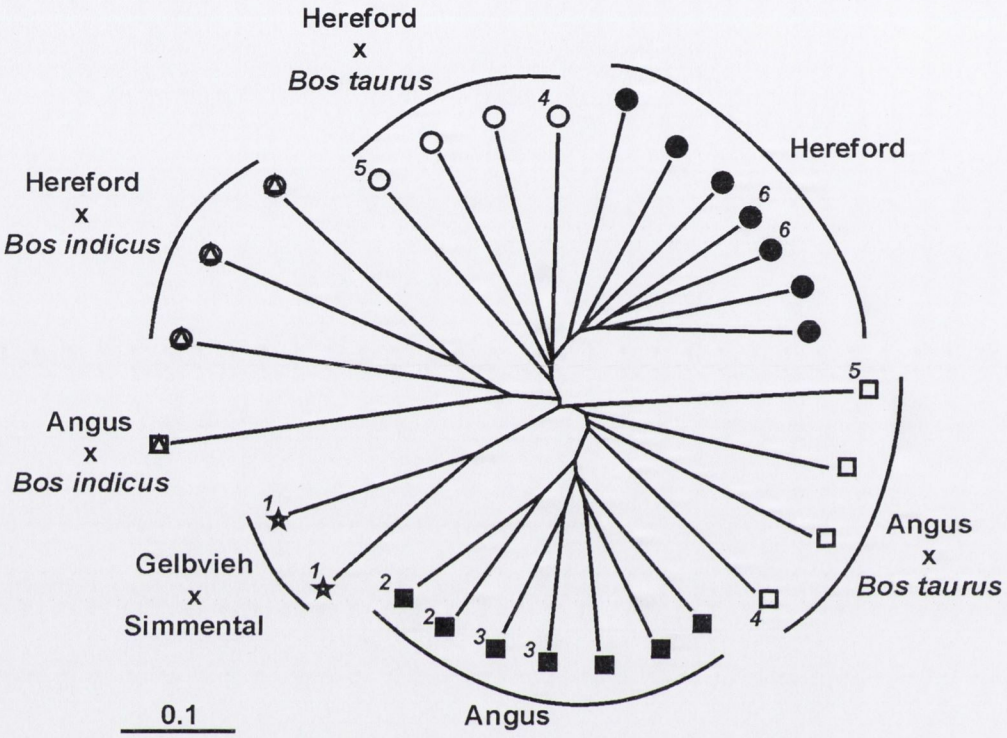


Figure 2.1 Neighbor joining tree constructed from a distance matrix based on allele sharing using the Neighbor application of Phylip (Felsenstein 1989).

Animals with Angus parentage are indicated with squares, while those with Hereford parentage are indicated with circles. Purebred animals are indicated by filled symbols and crossbred animals are indicated by open symbols. *Bos taurus-Bos indicus* crossbred animals are indicated with a triangle inside the outer symbol. Half-sib pairs are numbered.

2.3.1 Dendogram

Figure 2.1 shows a neighbour-joining tree of the 28 parental cattle based on a genome-wide allele-sharing matrix. This tree supports the known ancestral relationships: with the pure Angus and pure Hereford animals forming two respective clusters with the next closest animals to each being the Angus and Hereford crossbred animals. The longest branches lead to the *Bos indicus-Bos taurus* crossbreds. The Hereford pair of half sibs and the two Angus pairs of half sibs can be identified within the Hereford and Angus clusters (Figure 2.1) by slightly reduced branch lengths to their connecting nodes.

MultiDimensional scaling (Figure 2.2) of the same distance matrix highlights the greater separation between *Bos taurus* X *Bos indicus* cross bred animals and *Bos taurus* animals in dimension 1 (left to right) and a smaller separation between Angus and Hereford animals in dimension 2 (top to bottom).

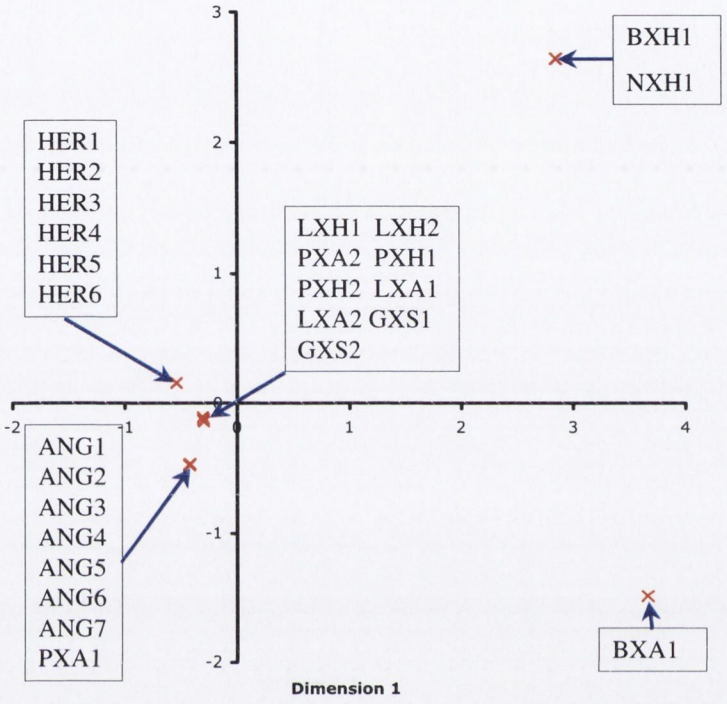


Figure 2.2 Multidimenstional scaling (MDS) showing the distance between samples constructed from the allele sharing distance matrix.

The ID codes are as follows: HER or H, Hereford; ANG or A, Angus; G, Gelbvieh; S, Simmental; L, Longhorn; P, Piedmontese; B, Brahman and N, Nellore. Brahman and Nellore are *Bos indicus* breeds and the remaining breeds are *Bos taurus* breeds.

The time since the divergence of *Bos taurus* and *Bos indicus* has been estimated to be in the order of hundreds of thousands of years before present from molecular data (Bradley et al. 1996; MacHugh et al. 1997), significantly predating the domestication of cattle approximately 10,000 years ago. Based on this known divergence and the data here, the *Bos indicus-Bos taurus* crossbred animals were omitted as known outliers from subsequent analyses of within-population diversity.

2.3.2 Analysis of microsatellite heterozygosity

In order to investigate the range of genetic diversity values for each microsatellite locus, the heterozygosity levels, which were calculated from allele frequency in the *Bos taurus* sample population assuming the IAM, were plotted against the number of alleles observed at that locus (Figure 2.3). Estimated 95% confidence intervals for the Ewens-

Watterson expected heterozygosity under the assumption of normality are given as a guide to assessing the deviation of outliers. When the observed heterozygosity was compared to the expected heterozygosity, 78 marker scores fell below the estimated 95% confidence interval and no loci fell above the 95% confidence interval.

An alternative representation of the genome-wide deviations of observed heterozygosity is a plot of the distribution of the standardised difference between observed and expected heterozygosity (DH/sd), obtained by dividing the difference between the observed and expected heterozygosity, which is computed assuming the IAM, by the standard deviation of the expected heterozygosity. The empirical distribution of genome-wide values of DH/sd for all polymorphic microsatellites is negatively skewed (Figure 2.4). A similar finding has been observed in human microsatellite data (Mattiangeli et al. 2006; Payseur et al. 2002). The same pattern was observed when the distribution was constructed using the TPM, the SMM or the empirical DH/sd.

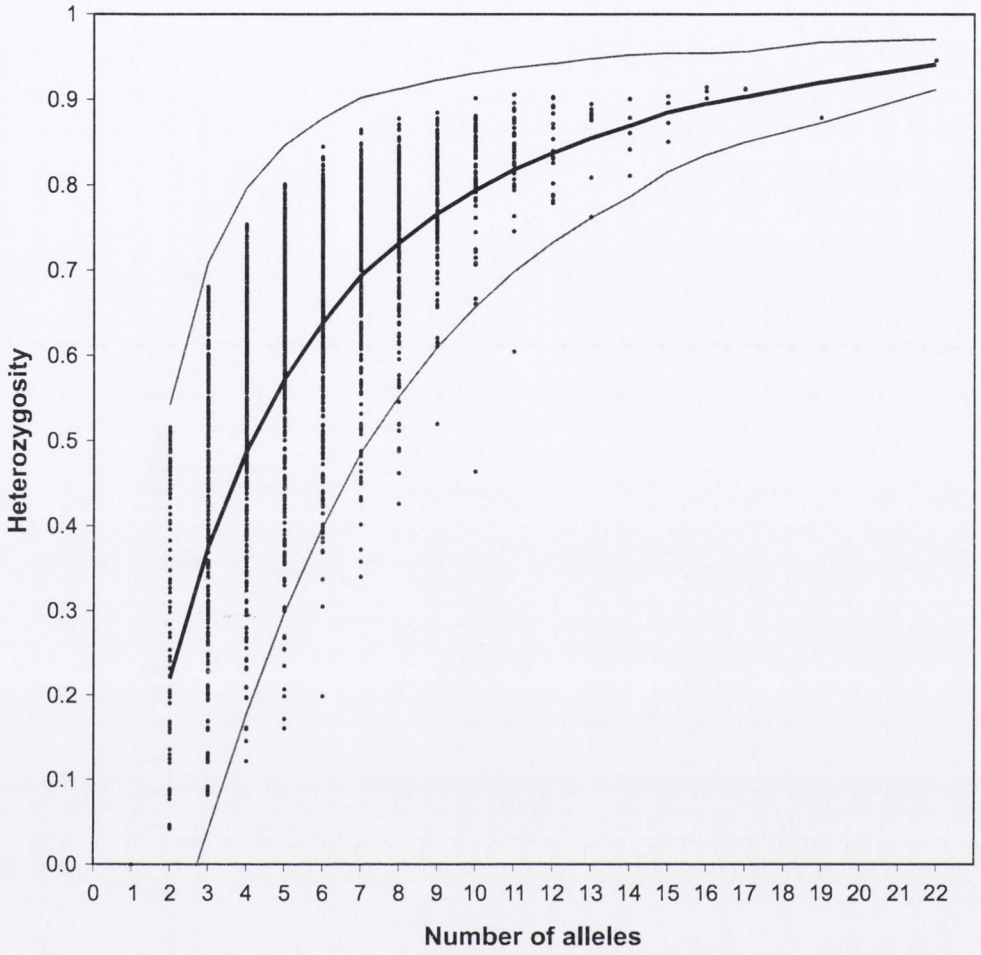


Figure 2.3 Departures from expected heterozygosity.

Observed heterozygosity (dot) for each polymorphic microsatellite plotted against number of alleles per locus. The approximate 95% confidence intervals (thin lines) for a locus with *n* alleles was estimated as the average expected heterozygosity (thick line) across all loci with *n* alleles ±1.96 times the average standard deviation (sd) of all loci for *n* alleles. The IAM was assumed. The positions of the loci with the most negative DH/sd values are indicated.

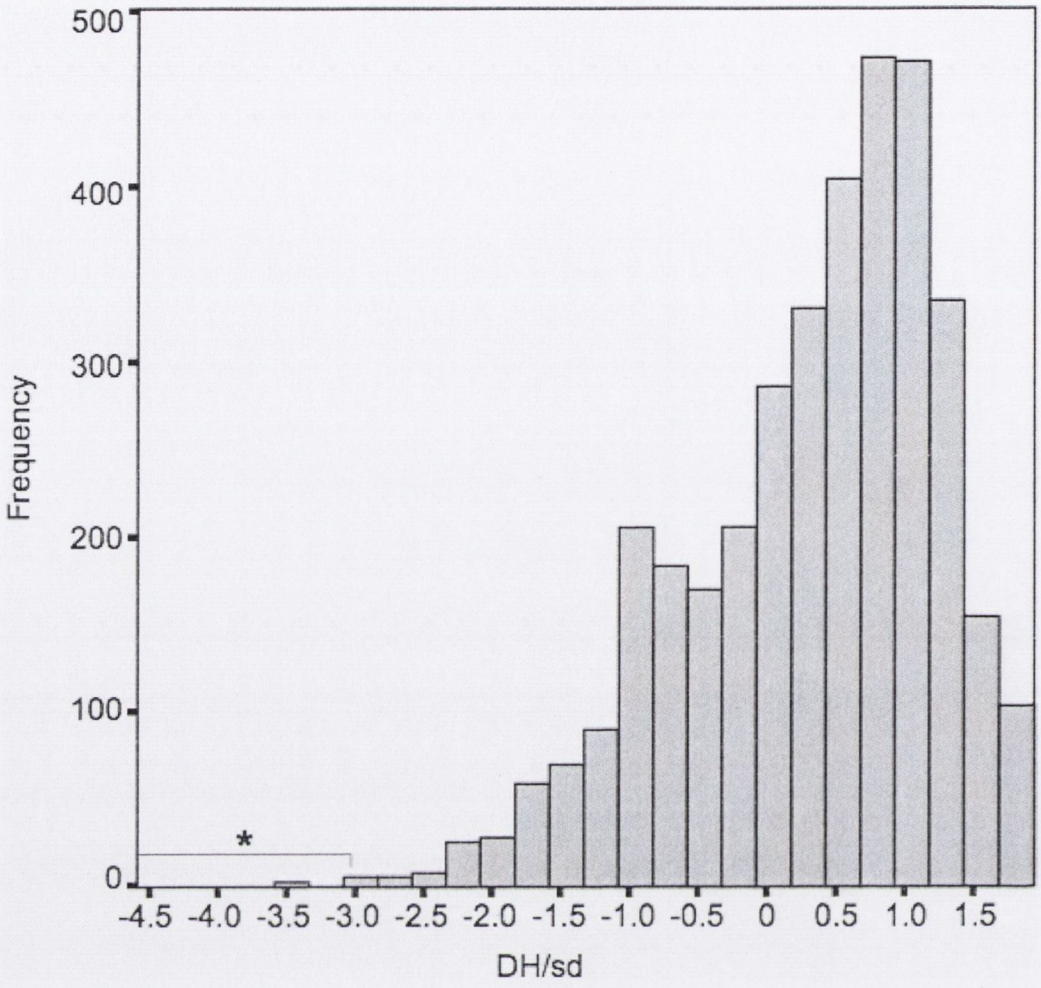


Figure 2.4 Genome wide distribution of DH/sd

Genome-wide distribution of the standardised difference between observed and expected heterozygosity for 3672 polymorphic microsatellites, assuming the IAM. The nine most extreme loci in the distribution are indicated by an asterisk.

2.3.3 Genes in the region of outlying markers

Nine of the ten markers at the negative end of the DH/sd distribution (which is suggestive of positive selection) are in the extreme tail of the distribution regardless of which of the three mutational models is assumed, suggesting that their position in the tail of the distribution is robust. When the empirical DH/sd is considered, these nine loci remain among the 15 most negative loci in the distribution (Table 2.2).

Table 2.2 Microsatellite loci at the negative tail of the DH/sd distribution.

Locus	IAM		TPM		SMM		Empirical	
DIK2443	1	-4.607	1	-8.270	1	-14.188	1	-5.128
BMS903	2	-4.164	2	-6.637	2	-10.946	2	-4.421
MNB-123	3	-3.499	3	-5.747	5	-9.460	5	-3.702
DIK036	4	-3.494	7	-5.316	9	-8.828	6	-3.549
DIK4523	5	-3.464	4	-5.488	4	-9.603	3	-4.421
INRA180	6	-3.325	9	-5.219	3	-10.080	4	-4.318
SRC240	7	-3.119	5	-5.484	10	-8.773	11	-3.377
DIK4492	8	-3.057	8	-5.226	9	-8.790	13	-3.286
NRDIKM024	9	-3.032	10	-5.062	6	-9.381	15	-3.242

For each model, the rank of the microsatellite (1 being most negative) and the value of DH/sd is given. These nine loci were chosen as being robust in the tail of the distribution as they were present in the tail regardless of what model was assumed.

Table 2.3 Map positions of loci at the negative end of the distribution.

Marker	Chromosome	Locus position
DIK2443	22	37866539 - 37866539,
BMS903	X	11786740-11786740
MNB123	29	27930303- 27930949
DIK036	15	40118055- 40118256
DIK4523		Could not be mapped to the genome assembly
INRA180	8	35796793- 35797239
SRC240	5	104194318- 104194554
DIK4492	22	6978913- 6979002
NRDIKM024	20	40177187- 40177425

The map positions of loci at the negative end of the distribution as determined by BLAST against the Ensembl assembly of the bovine genome (release 45, Btau_3.1).

Known and predicted genes within a 1Mb window centred on each of the nine loci were identified from the Ensembl release 45, Btau 3.1 (Hubbard et al. 2007) (Table 2.4) using

BioMart. The location of each gene and the distance to the microsatellite in bp was determined. Where no description was available for the bovine gene, the description of the human orthologue, as predicted by Ensembl, was used. In two cases no description was given for the putative bovine gene, and no human orthologue was predicted. In both of these cases, however, a murine orthologue was predicted and the description of this gene was recorded. For several genes there was no description and no predicted orthologues.

DIK036, lies less than 8kb downstream of an olfactory receptor (OR) gene cluster on BTA15. Homology to an olfactory receptor has simply been indicated by 'OR' due to inconsistencies with nomenclature and to the nature of this exceptionally large gene family. Order-specific gains and losses of OR genes have been shown to be frequent and the OR gene repertoire appears to be quite fluid on an evolutionary timescale (Niimura and Nei 2007). There is evidence to suggest that intact OR genes are under weak positive selection in humans (Gilad et al. 2003; Gilad et al. 2000)

Table 2.4 Genes within ± 0.5Mb of markers at the negative end of the distribution of DH/sd.

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
DIK2443	37353879 to 37531614	334925	synaptoporin [Source:RefSeq_peptide;Acc:NP_001039971]	ENSBTAG00000000265		SYNPR
	37540457 to 37541257	325282		ENSBTAG00000032868	pseudogene	
	37610965 to 37611878	254661		ENSBTAG00000030074		
	38009748 to 38401241	142980	Ca2+-dependent secretion activator [Source:RefSeq_peptide;Acc:NP_001069488]	ENSBTAG00000005321		CADPS
BMS903	11398939 to 11399268	387472		ENSBTAG00000034418		
	11425087 to 11426220	360520		ENSBTAG00000004363	Actin-related protein T1 (ARP-T1). [Source:Uniprot/SWISSPROT;Acc:Q8TDG2] [Target %id: 80; Query %id: 80]	ACTRT1
	12154655 to 12155691	367738		ENSBTAG00000025972	Histone-lysine N-methyltransferase SETMAR (EC 2.1.1.43) (SET domain and mariner transposase fusion gene-containing protein) (Metnase) (Hsmar1) [Includes: Histone-lysine N-methyltransferase; Mariner transposase Hsmar1]. [Source:Uniprot/SWISSPROT;Acc:Q53H47] [Target %id: 29; Query %id: 62]	SETMAR
MNB-123	27593123 to 27655298	275005	Friend leukemia virus integration 1 [Source:RefSeq_peptide;Acc:NP_001039763]	ENSBTAG00000008283		FLI1 (mouse)
	27685482 to 27686621	243682		ENSBTAG00000000008	ATP-sensitive inward rectifier potassium channel 1 (Potassium channel, inwardly rectifying subfamily J member 1) (ATP-regulated potassium channel ROM-K) (Kir1.1). [Source:Uniprot/SWISSPROT;Acc:P48048] [Target %id: 92; Query %id: 94]	KCNJ1
	27753075 to 27758146	172157		ENSBTAG00000006902	G protein-activated inward rectifier potassium channel 4 (GIRK4) (Potassium channel, inwardly rectifying subfamily J member 5) (Inward rectifier K(+) channel Kir3.4) (Heart KATP channel) (KATP-1) (Cardiac inward rectifier) (CIR). [Source:Uniprot/SWISSPROT;Acc:P48544] [Target %id: 93; Query %id: 93]	KCNJ5

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
MNB-123 (cntd.)	27788796 to 27789548	140755		ENSBTAG00000015903	pseudogene	
	27802544 to 27921402	8901		ENSBTAG00000015905	Rho GTPase-activating protein [Source: RefSeq_peptide (NP_055530)] [Target %id: 80; Query %id: 80]	
	28338688 to 28338915	407739		ENSBTAG00000033306		
	28391017 to 28391235	460068		ENSBTAG00000033305	none	
	28394958 to 28423838	464009		ENSBTAG00000003176	Junctional adhesion molecule C precursor (JAM-C) (Junctional adhesion molecule 3) (JAM-3) (JAM-2). [Source:Uniprot/SWISSPROT;Acc:Q9BX67][Target %id: 73; Query %id: 91]	JAM3
DIK036	39621359 to 39622309	495746		ENSBTAG00000027658	Olfactory receptor 52H1. [Source:Uniprot/SWISSPROT;Acc:Q8NGJ2] [Target %id: 85; Query %id: 86]	OR*
	39636085 to 39636654	481401	hypothetical protein LOC768030 [Source:RefSeq_peptide;Acc:NP_001070557]	ENSBTAG00000032226		OR*
	39648697 to 39649745	468310	ubiquilin-like [Source:RefSeq_peptide;Acc:NP_659490]	ENSBTAG00000005671		UBQLNL
	39666913 to 39669184	448871	*Bos taurus* ubiquilin 3 (UBQLN3), mRNA. [Source:RefSeq_dna;Acc:NM_001038583]	ENSBTAG00000025266		UBQLN3
	39684896 to 39686798	431257	LOC520778 protein (Fragment). [Source:Uniprot/SPTREMBL;Acc:Q32LK0]	ENSBTAG00000017647		
	39691386 to 39692369	425686		ENSBTAG00000032224	Olfactory receptor 52D1 (Olfactory receptor OR11-43) (Odorant receptor HOR5'beta14). [Source:Uniprot/SWISSPROT;Acc:Q9H346] [Target %id: 87; Query %id: 85]	OR*
	39698510 to 39699466	418589		ENSBTAG00000004707	Olfactory receptor 52D1 (Olfactory receptor OR11-43) (Odorant receptor HOR5'beta14). [Source:Uniprot/SWISSPROT;Acc:Q9H346] [Target %id: 85; Query %id: 85]	OR*
	39740464 to 39741405	376650		ENSBTAG00000018182		OR*

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
DIK036 (cntd.)	39758182 to 39759126	358929		ENSBTAG00000027653	Olfactory receptor 51I2 (Olfactory receptor OR11-38) (Odorant receptor HOR5'beta12). [Source:Uniprot/SWISSPROT;Acc:Q9H344] [Target %id: 89; Query %id: 88]	OR*
	39768208 to 39769158	348897		ENSBTAG00000027651	Olfactory receptor 51I2 (Olfactory receptor OR11-38) (Odorant receptor HOR5'beta12). [Source:Uniprot/SWISSPROT;Acc:Q9H344] [Target %id: 88; Query %id: 87]	OR*
	39790185 to 39791120	326935		ENSBTAG00000003582	Olfactory receptor 51I1 (Olfactory receptor OR11-39) (Odorant receptor HOR5'beta11). [Source:Uniprot/SWISSPROT;Acc:Q9H343] [Target %id: 85; Query %id: 86]	OR*
	39830564 to 39831505	286550		ENSBTAG00000019544	olfactory receptor 633 [Source:MarkerSymbol;Acc:MGI:3030467] [Target %id: 85; Query %id: 84] ***MOUSE***	OR*
	39842646 to 39847023	271032	Olfactory receptor 51M1 (Olfactory receptor OR11-40) (Odorant receptor HOR5'beta7). [Source:Uniprot/SWISSPROT;Acc:Q9H341]	ENSBTAG00000019549		OR*
	39943665 to 40008041	110014	Hemoglobin subunit beta (Hemoglobin beta chain) (Beta-globin). [Source:Uniprot/SWISSPROT;Acc:P02070]	ENSBTAG00000032216		HBB_BOVIN
	39953005 to 39954414	163641	Hemoglobin subunit epsilon-4 (Hemoglobin epsilon-4 chain) (Epsilon-4- globin). [Source:Uniprot/SWISSPROT;Acc:P06643]	ENSBTAG00000027648		HBE4_BOVIN
	39995013 to 39996606	121449		ENSBTAG00000032215	Hemoglobin subunit gamma-2 (Hemoglobin gamma-2 chain) (Gamma-2-globin) (Hemoglobin gamma-G chain) (Hb F Ggamma). [Source:Uniprot/SWISSPROT;Acc:P69892][Target %id: 61; Query %id: 65]	HBG2
	40028926 to 40030538	87517	Hemoglobin fetal subunit beta (Hemoglobin beta chain, fetal) (Beta- globin, fetal) (Hemoglobin gamma chain). [Source:Uniprot/SWISSPROT;Acc:P02081]	ENSBTAG00000032213		HBBF_BOVIN
	40043098 to 40044710	73345	Hemoglobin fetal subunit beta (Hemoglobin beta chain, fetal) (Beta- globin, fetal) (Hemoglobin gamma chain). [Source:Uniprot/SWISSPROT;Acc:P02081]	ENSBTAG00000032210		HBBF_BOVIN

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
DIK036 (cntd.)	40073236 to 40074099	43956		ENSBTAG00000027643		
	40080554 to 40081492	36563		ENSBTAG00000032208	Olfactory receptor 52A1 (Olfactory receptor OR11-319) (HPFH1OR) (Odorant receptor HOR3'beta4). [Source:Uniprot/SWISSPROT;Acc:Q9UKL2] [Target %id: 82; Query %id: 82]	OR*
	40126217 to 40127187	7961		ENSBTAG00000032205	Olfactory receptor 52A5 (Olfactory receptor OR11-33) (Odorant receptor HOR3'beta5). [Source:Uniprot/SWISSPROT;Acc:Q9H2C5] [Target %id: 83; Query %id: 83]	OR*
	40147537 to 40148472	29281		ENSBTAG00000032204	Olfactory receptor 52A1 (Olfactory receptor OR11-319) (HPFH1OR) (Odorant receptor HOR3'beta4). [Source:Uniprot/SWISSPROT;Acc:Q9UKL2] [Target %id: 83; Query %id: 83]	OR*
	40173889 to 40174833	55633		ENSBTAG00000032202	Olfactory receptor 52A1 (Olfactory receptor OR11-319) (HPFH1OR) (Odorant receptor HOR3'beta4). [Source:Uniprot/SWISSPROT;Acc:Q9UKL2][Target %id: 83; Query %id: 83]	OR*
	40186502 to 40187446	68246		ENSBTAG00000032201	Olfactory receptor 52A1 (Olfactory receptor OR11-319) (HPFH1OR) (Odorant receptor HOR3'beta4). [Source:Uniprot/SWISSPROT;Acc:Q9UKL2] [Target %id: 83; Query %id: 82]	OR*
	40208763 to 40266727	90507	ZCSL3 protein. [Source:Uniprot/SPTREMBL;Acc:Q0VBY7]	ENSBTAG00000032200	DPH4 homolog (CSL-type zinc finger-containing protein 3). [Source:Uniprot/SWISSPROT;Acc:Q6P3W2] [Target %id: 48; Query %id: 86]	DPH4
	40306085 to 40478853	187829	hypothetical protein LOC614139 [Source:RefSeq_peptide;Acc:NP_001033773]	ENSBTAG00000023831	Elongator complex protein 4 (hELP4) (PAX6 neighbor gene protein). [Source:Uniprot/SWISSPROT;Acc:Q96EB1] [Target %id: 35; Query %id: 74]	ELP4
	40485287 to 40513507	367031	Paired box protein Pax-6 (Oculorhombin). [Source:Uniprot/SWISSPROT;Acc:Q1LZF1]	ENSBTAG00000004561		PAX6_BOVIN

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
INRA180	35326181 to 35327089	469704		ENSBTAG00000033402	pseudogene	
	35451054 to 36020166	222927		ENSBTAG00000010178	Receptor-type tyrosine-protein phosphatase delta precursor (EC 3.1.3.48) (Protein-tyrosine phosphatase delta) (R-PTP-delta). [Source:Uniprot/SWISSPROT;Acc:P23468] [Target %id: 98; Query %id: 98]	PTPRD
SRC240	103475634 to 103727204	467114		ENSBTAG00000030357		
	103798508 to 103800304	394014	potassium voltage-gated channel shaker-related subfamily 5 [Source:RefSeq_peptide;Acc:NP_001015552]	ENSBTAG00000007553	Potassium voltage-gated channel subfamily A member 5 (Voltage-gated potassium channel subunit Kv1.5) (HK2) (HPCN1). [Source:Uniprot/SWISSPROT;Acc:P22460] [Target %id: 79; Query %id: 81]	KCNA5
	103822178 to 103828315	366003		ENSBTAG00000026522		
	103923048 to 103924535	269783		ENSBTAG00000016646	Potassium voltage-gated channel subfamily A member 1 (Voltage-gated potassium channel subunit Kv1.1) (HUKI) (HBK1). [Source:Uniprot/SWISSPROT;Acc:Q09470] [Target %id: 96; Query %id: 96]	KCNA1
	104022308 to 104023894	170424		ENSBTAG00000030352	Potassium voltage-gated channel subfamily A member 6 (Voltage-gated potassium channel subunit Kv1.6) (HBK2). [Source:Uniprot/SWISSPROT;Acc:P17658] [Target %id: 94; Query %id: 94]	KCNA6
	104056660 to 104095035	99283	LOC540652 protein (Fragment). [Source:Uniprot/SPTREMBL;Acc:A0JNJ8]	ENSBTAG00000017378	Probable polypeptide N-acetylgalactosaminyltransferase 8 (EC 2.4.1.41) (Protein-UDP acetylgalactosaminyltransferase 8) (UDP- GalNAc:polypeptide N-acetylgalactosaminyltransferase 8) (Polypeptide GalNAc transferase 8) (GalNAc-T8) (pp-GaNTase 8). [Source:Uniprot/SWISSPROT;Acc:Q9NY28][Target %id: 63; Query %id: 80]	GALNT8

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
SRC240 (cntd.)	104126685 to 104148510	45808	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial precursor (EC 1.6.5.3) (EC 1.6.99.3) (NADH-ubiquinone oxidoreductase 39 kDa subunit) (Complex I-39kD) (CI-39kD). [Source:Uniprot/SWISSPROT;Acc:P34943]	ENSBTAG00000005465		NDUFA9
	104161125 to 104176322	17996	A-kinase anchor protein 3 (Protein kinase A-anchoring protein 3) (PRKA3) (A-kinase anchor protein 110 kDa) (AKAP 110) (Fragment). [Source:Uniprot/SWISSPROT;Acc:O77797]	ENSBTAG00000001020		AKAP3
	104181069 to 104248366	13249	hypothetical protein LOC615937 [Source:RefSeq_peptide;Acc:NP_001069871]	ENSBTAG00000007556	Dual specificity tyrosine-phosphorylation-regulated kinase 4 (EC 2.7.12.1). [Source:Uniprot/SWISSPROT;Acc:Q9NR20] [Target %id: 81; Query %id: 66]	DYRK4
	104248745 to 104288522	54191	hypothetical protein LOC515243 [Source:RefSeq_peptide;Acc:NP_001069172]	ENSBTAG00000009227		
	104319843 to 104331856	125289		ENSBTAG00000006800	Fibroblast growth factor 6 precursor (FGF-6) (HBGF-6) (HST-2). [Source:Uniprot/SWISSPROT;Acc:P10767] [Target %id: 91; Query %id: 91]	FGF6
	104370114 to 104378687	175560		ENSBTAG00000030343	Fibroblast growth factor 23 precursor (FGF-23) (Tumor-derived hypophosphatemia-inducing factor). [Source:Uniprot/SWISSPROT;Acc:Q9GZV9] [Target %id: 76; Query %id: 78]	FGF28
	104385007 to 104399976	190453	hypothetical protein LOC615392 [Source:RefSeq_peptide;Acc:NP_001069838]	ENSBTAG00000016650		
	104415749 to 104439011	221195	hypothetical protein LOC615414 [Source:RefSeq_peptide;Acc:NP_001069840]	ENSBTAG00000016649	G1/S-specific cyclin-D2. [Source:Uniprot/SWISSPROT;Acc:P30279] [Target %id: 94; Query %id: 93]	CCND2
DIK4492	6993950 to 6994578	14948		ENSBTAG00000034604		
NRDIKM024	39564397 to 39816720	360467		ENSBTAG00000012558	ADAMTS-12 precursor (EC 3.4.24.-) (A disintegrin and metalloproteinase with thrombospondin motifs 12) (ADAM-TS 12) (ADAM-TS12). [Source:Uniprot/SWISSPROT;Acc:P58397] [Target %id: 75; Query %id: 83]	ADAMTS12

Marker	Gene position	Distance (bp)	Description (bovine gene)	Ensembl ID	Description (human)	Gene Symbol
NRDIKM024 (cntd.)	39719240 to 39929767	247420		ENSBTAG00000032901		
	39762184 to 40014698	162489	Atrial natriuretic peptide clearance receptor precursor (ANP-C) (ANPRC) (NPR-C) (Atrial natriuretic peptide C-type receptor). [Source:Uniprot/SWISSPROT;Acc:P10730]	ENSBTAG00000007665		NPR3
	40095162 to 40114685	62502		ENSBTAG00000021965	SUB1 homolog (S. cerevisiae) [Source:MarkerSymbol;Acc:MGI:104811] [Target %id: 93; Query %id: 93] ***MOUSE***	Sub1
	40264604 to 40325429	87179		ENSBTAG00000008414	zinc finger RNA binding protein [Source: RefSeq_peptide (NP_057191)] [Target %id: 94; Query %id: 97]	ZFR
	40351956 to 40448055	174531		ENSBTAG00000012278	myotubularin related protein 12 [Source: RefSeq_peptide (NP_001035536)] [Target %id: 79; Query %id: 77]	MTMR12
	40361957 to 40381698	184532		ENSBTAG00000032855	F-box/LRR-repeat protein 18 (F-box and leucine-rich repeat protein 18). [Source:Uniprot/SWISSPROT;Acc:Q96ME1] [Target %id: 71; Query %id: 81]	FBXL18
	40490830 to 40498652	313405		ENSBTAG00000032853	Heme oxygenase 1 (EC 1.14.99.3) (HO-1). [Source:Uniprot/SWISSPROT;Acc:P09601] [Target %id: 59; Query %id: 54]	HMOX1
	40498239 to 40550184	320814	golgi phosphoprotein 3 [Source:RefSeq_peptide;Acc:NP_001069416]	ENSBTAG00000032848		GOLPH3
	40562946 to 40816589	385521		ENSBTAG00000014455	PDZ domain-containing protein 2 (PDZ domain-containing protein 3) (Activated in prostate cancer protein). [Source:Uniprot/SWISSPROT;Acc:O15018] [Target %id: 66; Query %id: 71]	PDZD2

Gene position and distance from the microsatellite are indicated in bp. Where available, the description of the bovine gene is given. Where no description was available for the bovine gene, the description of human orthologue was used. In two instances (**) no human orthologue was reported, and the description of the murine homologue was used. a) Numerous genes with homology to olfactory receptors (OR) were identified. In the gene symbol these have been identified as OR*.

2.3.4 Correlation between Lab data and USDA USMARC data

77 microsatellite markers that were included in the USDA USMARC data set had previously been genotyped in Charolais and Friesian (Park 2002). DH/sd values for these data were calculated and were compared to the values calculated from the USDA USMARC cattle reference family population by Pearson correlation. Figure 2.5 shows the DH/sd values for a) the USMARC sample plotted against Charolais, b) the USMARC sample plotted against Friesian and c) Charolais plotted against Friesian. The Pearson correlation and significance are indicated on each figure. The DH/sd values used here were calculated assuming the IAM, but similar results were obtained when the SMM was assumed (not shown). The correlation between the USMARC data and the Charolais was close, but not significant, and the correlation between USMARC and Friesian, and Charolais and Friesian were significant. MacHugh et al. (1994) showed that the genetic distance (Nei's standard genetic distance) between Hereford and Charolais (0.077) and Angus and Charolais (0.123), the breeds predominant in the USMARC cattle reference families, was greater than the distance between either and Friesian (0.051 and 0.102 respectively). This may explain the lower correlation between the USMARC data and Charolais data.

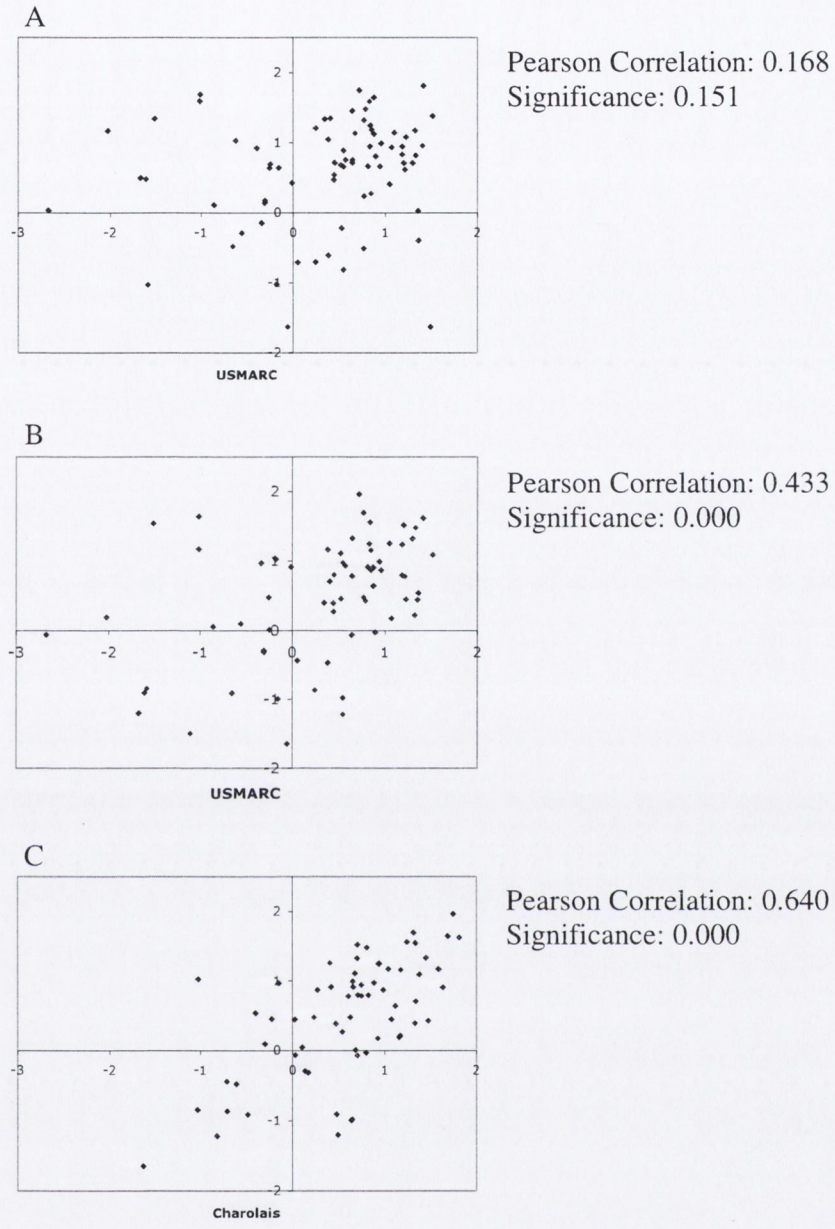


Figure 2.5 Correlation between DH/sd of 77 markers genotyped in three populations.

A) Correlation between DH/sd in the 24 *Bos taurus* parents of the USDA USMARC cattle reference families (USMARC) on x-axis and a Charolais sample population (Park 2002); B) Correlation between DH/sd in the 24 *Bos taurus* parents of the USDA USMARC cattle reference families (USMARC) on x-axis and a Friesian sample population (Park 2002); C) Correlation between DH/sd in the Friesian and Charolais populations. The Pearson correlation and significance (two tailed) are indicated for each comparison.

2.4 Discussion

2.4.1 Sample population

By comparison with similar studies, the present work has the advantage of reasonably dense marker coverage at over 3800 markers covering all the autosomes and the X chromosome. However, the sample population from which the genotype data were obtained is imperfect. The sample size is low, having only 24 *Bos taurus* animals. Six pairs of half sibs are present in the sample and more critically, the sample does not represent a single breed or a random sample across a wider population. Instead, the sample is predominantly composed of Angus and Hereford. With this in mind, and given the availability of such a large dataset, the analysis proceeded.

2.4.2 Observed and Expected heterozygosity

The Ewens-Watterson approach allows an assessment of allele frequency spectra at marker loci which may highlight markers showing patterns of reduced diversity; such as should emerge as a result of positive selection of a linked advantageous genetic variant.

A number of loci were identified that showed deviations from expected heterozygosity (Figure 2.3). All of the loci that fell outside the 95% confidence intervals exhibited a deficit of heterozygosity. No loci were identified showing an excess of heterozygotes suggesting that the inclusion of cross-bred animals is not of major importance in the sample. All of the deviations showed a deficit of heterozygosity, however, the number of loci falling outside the 95% confidence interval is no greater than would be expected. 78 loci out of 3672 (3765 loci less the 93 loci that were monomorphic in *Bos taurus*) or 2.1% fell below the 95% confidence interval.

The compilation here of a genome-wide series of such values enables the comparison of individual marker results versus an empirical distribution of DH/sd test results. This aids in distinguishing localised effects that may be due to selection from genome-wide imprints of demographic processes. The distribution of genome-wide DH/sd values is similar in shape to that of Mattiangeli et al. (2006), who analysed microsatellite diversity in an Irish human population. One minor difference is that more positive values were observed in the data represented here. This seems to be a general trend in cattle (see Figure 3.10 and Figure 3.11 in Chapter 3). A second point to note is a small peak in the distribution at -1.

The DH/sd measurement of microsatellite variability has previously been reported as evidence supporting the action of positive selection in the region downstream of *Notch* on the X chromosome of *Drosophila melanogaster*; confirming inference from other marker systems (Bauer DuMont and Aquadro 2005). Mattiangeli *et al.* have also used outliers on an empirical DH/sd distribution in a human population to support earlier evidence for a selective history at three human loci, including the cystic fibrosis gene (Akey et al. 2002; Mattiangeli et al. 2006).

In addition to considering the standard DH/sd values, we calculated an empirical DH/sd value for each locus. Based only on the observed data, it does not rely on assumptions about the mutation model of microsatellites, as the conventional DH/sd measure does. Similar findings were seen for both the empirical DH/sd and the conventional DH/sd, particularly when the IAM was assumed. The distributions were very similar in both shape and range, and overall, there was good agreement between the loci at the tail of one distribution and the loci at the tail of the other distribution. (Table 2.2). This observation offers some validation of the approach taken here and suggests that the position of these loci in the tails of the distribution is not an artefact resulting from the assumptions about the model of microsatellite evolution.

2.4.3 Genes near the negative end of the distribution of DH/sd

Lastly, one olfactory receptor gene cluster on BTA15 is proximal to the 4th ranked microsatellite (Table 2.4). Olfactory receptors are intriguing candidates for selective effects, perhaps due to pressure for odour recognition as a result of modified behaviour, *e.g.* foraging strategy, due to capture and herding. Analysis of stable isotopes in archaeological bone samples of wild and early domestic cattle from northern Europe suggest different diets. Data from Aurochs are consistent with forest foraging whereas early domesticates seem to have fed on open grassland (Noe-Nygaard et al. 2005).

These genes should not be considered as strong candidates but as more bovine genetic data becomes available, these regions can be revisited. Large scale SNP analysis emerging from the bovine genome project will be a significant advance in this field. Microsatellites and SNPs are correlated, reflecting a shared genealogical history, but correlation is weak, most likely due to different underlying mechanisms of mutation and mutation rates (Payseur and Cutter 2006). Combining microsatellite analyses with SNP analyses may reveal aspects of the evolutionary landscape of the genome that cannot be revealed by either alone. Furthermore, additional data may help resolve the question of

whether the empirical approach with large numbers of markers is sufficient to identify outliers regardless of population problems.

2.4.4 Marker redundancy

One test for a regional biological signal over marker-specific random noise is whether extreme Ewens-Watterson values are observed across neighbouring markers. If strongly negative values of DH/sd are the result of positive selection, one might expect to see congruent evidence of past selective sweeps within genomic regions, *i.e.* neighbouring loci to show similar deficits in observed heterozygosity. In order to investigate this, the loci at the tails of the DH/sd distribution were considered. Of the 100 loci falling at the negative end of the DH/sd distribution, six pairs of neighbouring loci (i.e., loci which are adjacent on the genetic map) were present. When 100 loci were chosen at random from the dataset, six or more pairs were found in fewer than 5% of 10,000 replications. The same pattern was not observed at the positive end of the distribution, with only two pairs in the top 100 loci (found in over 75% of random selections). However, sequence analysis of the microsatellite loci, where available in GenBank, and primer sequences revealed that of the six adjacent pairs of loci identified at the negative tail of the distribution, four included the same repeat region amplified with different primers and assigned different marker names. A further pair showed discordant map positions based on sequence comparison to the genome assembly.

Here, an apparent excess of similarly outlying scores in neighbouring outliers proved to be due to a number of instances where the same microsatellite sequence has been amplified using different primer sets. This is not surprising given that the bovine linkage map is a collaborative compilation of markers from diverse laboratory sources. The extent of marker redundancy in this dataset is unknown, and it is beyond the scope of the current project to investigate it thoroughly, but caution would be urged when selecting multiple microsatellites from the same region. Furthermore, a number of instances were noted where genotype differences were observed for markers that amplified the same microsatellite sequence with different primers. Whether these differences are due to genotyping error or flanking variation unique to specific primer pairs (or both) is unknown.

2.5 Conclusions

The population structure of the sample of animals available for this study can be justifiably criticised as being less than ideal. The deficit of heterozygotes – as indicated by departures from expected heterozygosity (Figure 2.3) and the tail in observed DH/sd distribution (Figure 2.4) - could be an effect due to the heterogeneous population. Human studies in more ideal populations, for example the Irish population (Mattiangeli et al. 2006) show the same shape of DH/sd distribution, however, and so it is unlikely that this effect is purely down to population structure. In addition to the undesirable heterogeneity, the number of animals in the population was too small to achieve the desired power to robustly detect specific signatures of selection. Despite this, the choice of mutational model does not greatly affect the outcome, either in terms of the overall shape of the data or the extreme outliers. Even if individual results should be treated with caution, therefore, the general trends are probably reliable.

Problems due to a sub-optimal population are an expected consequence of relying on pre-existing data. An unexpected problem is the marker redundancy discovered during the course of this study when examining neighbouring loci. It would be difficult to determine how many markers are affected by redundancy and/or inaccurate mapping to the genome; such inaccuracies are particularly likely in a genome that is still being completed and annotated such as is the case for cattle. These errors reduce confidence in results generated using them and so should be quantified or affected loci removed if this type of analysis is to be repeated successfully. Furthermore, the results presented here highlight the need for a larger, and preferably homogeneous, population.

3 Microsatellite Variation Around a Candidate Region for Selection

3.1 Introduction

3.1.1 Natural Killer cell receptors and the immune response

Natural Killer (NK) cells play an important role in the innate immune response, which is vital for eliminating infection in the early stages or containing the infection until the adaptive immune response can be activated. Characterised by large cytoplasmic granules, these cells of the lymphoid lineage are produced in the bone marrow and circulate in the blood in a partially activated state, enabling a swift response to infection. NK cells are primarily active against intracellular pathogens, including certain viruses, bacteria and protozoa, but are also active against certain tumour cells. Target cells are killed by apoptosis, which is induced by the release of perforin and granzymes contained within the cytoplasmic granules of the NK cell onto the surface of the bound target cell.

The cytotoxic activity of NK cells is regulated by a balance between signals from the two functional classes of NK receptors: activating receptors and inhibitory receptors. The functionality of these receptors is determined by the cytoplasmic region of the protein. Those having an inhibitory function have a long cytoplasmic tail containing an immunoreceptor tyrosine-based inhibition motif (ITIM), which becomes phosphorylated upon stimulation and transduces the inhibitory signal to the cell via an intracellular phosphate such as SHP1 or SHP2, blocking the cytotoxic activity of the cell. Those receptors having an activating function have shorter cytoplasmic tails, lack the ITIM and contain a positively charged residue in the transmembrane (TM) region. This positively charged residue associates with a negatively charged residue of an immunoreceptor tyrosine-based activation motif (ITAM) bearing adaptor molecule, such as DAP12, also known as TYROBP, for transduction of the activating signal (Campbell and Colonna 1999).

The ligands for inhibitory receptors are MHC class I molecules, and ligand binding inhibits the cytotoxic activity of NK cells. Cells expressing abnormal levels of MHC class I molecules, perhaps as a consequence of viral infection or transformation to tumour cell, are incapable of inhibiting NK cell activity and are killed. The ligands for activating receptors are largely unknown, but appear to include both MHC and non-MHC molecules (Trowsdale et al. 2001; Uhrberg 2005)

In addition to functional classification, the NK receptors can be divided into two families – the C-type lectin receptors and the immunoglobulin-like receptors – based on the structure of the extracellular region of the protein. In both the mouse and the human

genome, the majority of genes encoding these receptors are located in two clusters – the C-type lectin family in the Natural Killer Complex (NKC) on HSA12 or MMU6 and the immunoglobulin-like receptors in the Leukocyte Receptor Complex (LRC) on HSA19 and MMU7 (Figure 3.1). Emerging data from the sequencing of other mammalian genomes indicates that this clustering of receptors is conserved across mammalian species. However, the gene content within these clusters is extremely variable between mammalian lineages.

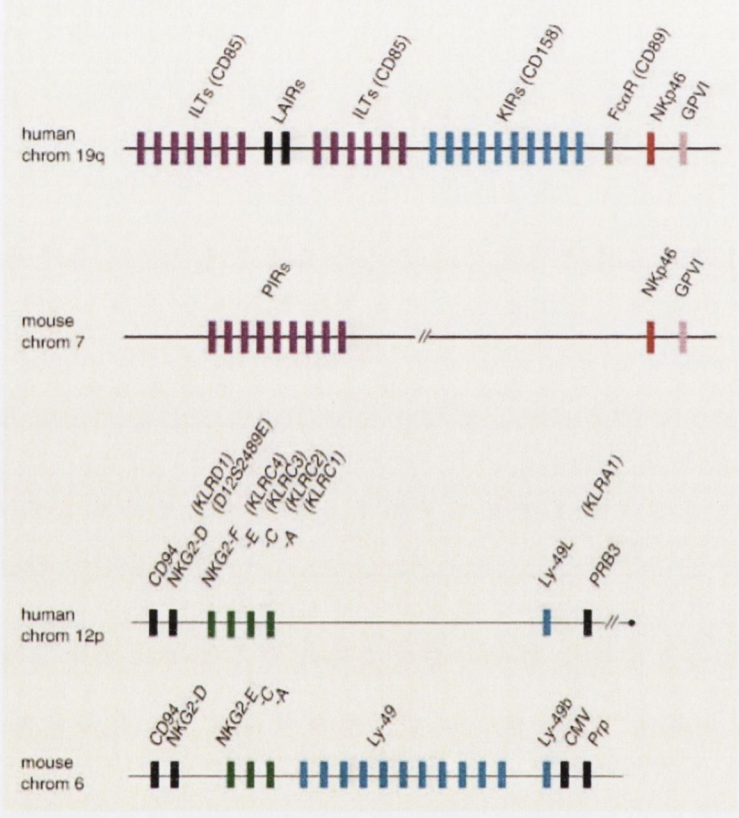


Figure 3.1 A schematic diagram illustrating the expansion of the KIR genes located in the LRC on human chromosome 19, and of the Ly49 genes located in the NKC on mouse chromosome 6. (Taken from Trowsdale et al. 2001)

3.1.1.1 KIR genes

The major class of NK cell receptors in human are the Killer cell Immunoglobulin-like Receptors (KIR), a subset of the Immunoglobulin-like receptors. The KIR genes are located within a 100kb region of the LRC on HSA19 and are arranged in a head-to-tail fashion, typically separated by only 1-2kb. In mouse and rat, a class of evolutionarily unrelated receptors, the Ly49 receptors, which are members of the C-type lectin family, is

the major class of NK cell receptors, carrying out the same function that the KIR genes carry out in humans.

The early thinking was that KIR genes had arisen along the branch leading to primates and that the major NK receptor class in other mammalian species was the Ly49s, as KIR genes had only been identified in primates and none had been identified in rodents. This is now known not to be the case. KIR genes have been identified or inferred to exist in mouse, rat, horse, cow, pig and hamster in addition to baboon, rhesus monkey, orangutan, gorilla, bonobo, chimpanzee and human (Kelley et al. 2005; Parham 2005; Volz et al. 2001).

Seventeen KIR genes, including two pseudogenes, and only a single Ly49 gene, which is also a pseudogene, have been have been identified in the human genome to date (Uhrberg 2005), while the mouse genome appears to contain 16 Ly49 genes and two atypical KIR genes, being both located on the X chromosome. Although evolutionarily unrelated, the KIR genes in humans and the Ly49 genes in mouse have undergone a similar expansion in order to carry out the same function, and this is achieved by a very similar signalling mechanism, providing a clear example of convergent evolution. Multiple KIR genes have also been identified in non-human primate species, including chimpanzee, gorilla, orangutan, Rhesus macaque and the African green monkey (Kelley et al. 2005). Analysis of retroelement sequences in the region of the KIR genes allowed the expansion of KIR genes in the human lineage to be dated to between 30-45 million years ago, some time after the mammalian radiation (Martin et al. 2000).

The HUGO nomenclature for KIR genes is made up of i) the number of Ig domains that the gene has (i.e., 2D or 3D), ii) the length of the cytoplasmic tail type (S indicating a short tail or L indicating a long tail) which therefore describes the functionality of the gene, and iii) an identifier number indicating the order the genes were named. Thus KIR2DS1 is a KIR gene containing two Ig domains, has a short cytoplasmic tail, and therefore is most likely an activating receptor, and was the first 2DS gene described in that species. It should be noted that this nomenclature system provides no information on the relationship between KIR genes from different species.

The Ig domains of KIR genes are designated D0, D1 and D2, from the N-terminal of the protein, and are each encoded by a separate exon. Most primate KIRs containing two Ig domains are of the type D1-D2 (Martin et al. 2000), however the two bovine KIRs containing two domains described to date are of the types D0-D1 and D0-D2, suggesting

that an ancestral KIR containing three Ig domains existed prior to the divergence of the bovine and primate lineages.

An investigation of copy number of Ly49 and KIR genes in the bovine genome by Southern blot analysis revealed that while Ly49 is present in a single copy, multiple KIR genes are present (McQueen et al. 2002). Analysis of the bovine Ly49 gene showed it to be more similar to the human Ly49 at the sequence level, but suggested that unlike the human Ly49, it is likely to be functional.

Molecular cloning and sequencing of four bovine KIR genes revealed that this group contain both activating and inhibitory receptors. KIR2DL1 and KIR3DL1 contain an ITIM in their cytoplasmic region, indicating inhibitory function, while KIR2DS1 and KIR3DS1 lack the ITIM but have an arginine containing motif in their TM region suggesting that these function as activating receptors.

Using a bovine/hamster radiation hybrid (RH) panel, KIR2DL1 and NKp46 were mapped to a region of BTA18, and were most closely linked to the marker TGLA132. Similarly, KLRD1 and KLRJ1 were mapped to a region of BTA5, and were most closely linked to the marker BM1658. These regions of BTA18 and BTA5 are homologous to HSA19 and HSA12, respectively, where the human LRC and NKC are located, suggesting a similar arrangement of bovine NK cell receptors in cattle. This was further supported by the mapping of the bovine FcaR to the same region on BTA18 as KIR2DL1 and NKp46 (Morton et al. 2004).

Phylogenetic analysis of two bovine KIR sequences (KIR2DL1 and KIR3DL1) revealed that they form a separate branch from primate KIR genes, indicating that the expansion of KIR genes in the bovine lineage was likely a separate but parallel event to that which occurred in the primate lineage (McQueen et al. 2002). Phylogenetic analysis of individual Ig domains of primate and bovine KIR genes further supported the theory of a parallel expansion of KIR genes in cattle and primates from a single ancestral gene, showing bovine Ig domains forming distinct clusters within each domain class (Figure 3.2) (Storset et al. 2003).

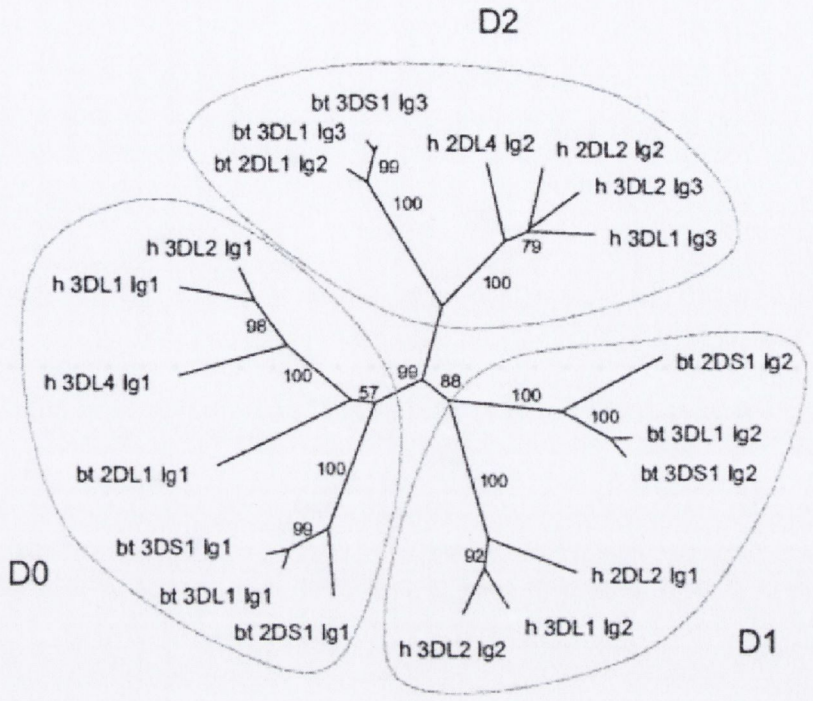


Figure 3.2 Phylogenetic tree constructed using amino acid sequences of individual Ig domains from bovine (bt) and selected primate (h) KIR genes.
(Taken from (Storset et al. 2003).)

Given the close functional relationship between MHC loci and KIR loci, and the patterns of diversity in human KIR genes it has been suggested that KIR genes may be subject to similar evolutionary forces as MHC loci (Martin et al. 2002; Single et al. 2007; Vilches and Parham 2002). The MHC is widely regarded as a classical example of balancing selection and it is thought that this force is responsible for maintaining the exceptionally high levels of polymorphism observed at MHC loci (Hedrick and Thomson 1983). Balancing selection has also been reported at MHC loci in ruminants (Park 2002; Paterson 1998) and high levels of diversity have been reported at MHC loci in an otherwise monomorphic mammal, the San Nicholas Island Fox (Aguilar et al. 2004).

Genetic variability at the KIR cluster in humans has been reported to be indicative of balancing selection (Norman et al. 2004). More recently, combined analysis of diversity at several KIR loci and their respective HLA ligands in global human populations has suggested that the KIR region is influenced by complex selective pressures differentially affecting activating and inhibitory KIR loci (Single et al. 2007). A strong negative correlation was observed between the presence of activating KIRs and their HLA ligands, wheras a weak positive correlation was observed between inhibitory KIRs and their HLA

ligands. Intriguingly, a negative correlation between the population frequency of activating KIRs and the distance from East Africa was also observed. This observation may be associated with the migration of human populations out of Africa.

3.2 Materials and Methods

3.2.1 DNA samples

A total of 254 samples from 6 populations representing cattle breeds from around the world were included in this study.

DNA samples from two *Bos indicus* populations, Ongole (n=30) and Nellore (n=27) were provided by Claude Gaillard, Institiut für Tierzucht, University of Berne, Switzerland. The Nellore samples were obtained in Brazil, and the Ongole samples were obtained from herds maintained at Visakhapatnam Government Dairy farm, Andhra Pradesh.

The two European *Bos taurus* populations were Charolais (n=33) and Friesian (n=40). For both groups, blood or semen samples were obtained from artificial insemination (AI) stations in Britain and Ireland, as described in MacHugh (1996) and Park (2001). Pedigree information was available for Charolais and Friesian animals. These records were consulted to ensure that closely related individuals were not sampled.

The two African *Bos taurus* populations included were Somba (n=61) and N'Dama (n=63) . The Somba samples were collected in northern Benin in collaboration with Dr. L. Gnaho, Benin, and N'Dama samples were collected in Boke, Boffe and Gaoul in Western Guinea, as described by Park (2002). Blood samples were collected and transported on ice back to the laboratory for DNA extraction. DNA was obtained from the whole blood samples by a salt extraction protocol as described by MacHugh (1996). Pedigree information was not available for Somba and N'Dama but by sampling no more than one tenth of any herd, and by consulting the owners and herders the chance of sampling related individuals was reduced.

3.2.2 Microsatellites

17 microsatellites located across an approximately 30cM region at the telomeric end of BTA18 were genotyped in the six bovine population samples. These microsatellites included 14 taken from public databases and three novel loci.

3.2.2.1 Publicly available microsatellites

Microsatellites and primer sequences were taken from the COMRAD database (http://www.projects.roslin.ac.uk/comrad) and the MARC USDA database (http://www.marc.usda.gov/genome/cattle/cattle.html). Microsatellites were selected on the basis of being in the region surrounding the KIR gene cluster on BTA18, and having no evidence of null alleles. The microsatellites are outlined in Table 3.2.

3.2.2.2 Novel microsatellites

In order to increase the resolution of the study, novel tandem repeats that could potentially be polymorphic were identified in the region around the LRC.

Nucleotide sequences for four bovine KIR genes, bKIR2DL1, bKIR2DS1, bKIR3DL1 and bKIR3DS1, were available in the NCBI database. In addition, sequences were available for the bovine FcaR and NKp46 genes. The human homologues of these genes are located at the telomeric end of the KIR cluster on HSA19. The nucleotide sequences were used to identify the UniProt accession number for each gene, where possible. The UniProt accession numbers were used to search the EST assembly browser (http://popgen.gen.tcd.ie/snps4) in order to identify EST contigs for these genes. A blastx against nr limited to *Bos taurus* was carried out with the consensus sequence of each EST contig. For those sequences with no bovine entry in UniProt, the human homologues were used.

Web BLAT was used to assign each EST contig to a bovine genome project scaffold. Tandem Repeat Finder (Benson 1999) was used to identify tandem dinucleotide repeats on these scaffolds (Lilian Lau). By choosing tandem repeats of highest number of repeated units, it was hoped to maximize the possibility of identifying polymorphic loci. By this method, four scaffolds were identified which contained tandem repeats. PCR primers were designed using Primer3 (Rozen and Skaletsky 2000) to amplify 5 of these regions, designated MS18-1 to MS18-5.

3.2.2.3 Genomic positions of microsatellites

In order to examine the microsatellite variation across the region, it is necessary to establish the order of markers relative to each other. Information on marker position and order along the chromosome was available in cM for all those markers taken from the MARC database (Table 3.2); the microsatellites obtained from the COMRAD database, KSCG and TGLA132, were ordered by radiation hybrid mapping and relative locations were available in the form of a map showing order along the chromosome. The novel microsatellites had neither centiMorgan positions nor were they included in the radiation hybrid mapping.

Marker positions were determined using the *in silico* PCR application available at the UCSC Genome browser (http://genome.ucsc.edu/). The August 2006 release (Btau_3.1) was used. *In silico* PCR allows the user to search the bovine genome with a forward and

a reverse primer sequence and returns any hits that match the user specified parameters (i.e. primer sequences, product length).

In the case of IDVGA-55, *in silico* PCR failed to return a match in the bovine genome. The position of this microsatellite was obtained by carrying out a BLAT search using the GenBank sequence (X85071).

3.2.3 PCR amplification and genotyping

PCR primers to amplify each microsatellite were ordered with the fluorescent dye on the 5' end of each forward primer. The fluorescent probes used were HEX (green), TAMRA (yellow) and 6'FAM (blue) and modified primers were designed such that multiple markers with products of different sizes and different fluorescent probes could be run on a single acrylamide gel without risk of overlapping allele size ranges.

Where the expected size of the PCR product was unknown, an unmodified primer was first ordered, and the approximate size of the product was estimated by reference to a 100bp DNA ladder (Invitrogen) run out on a 2% agarose gel stained with ethidium bromide. A replacement forward primer, modified with a 5' fluorescent dye was then ordered to replace the unmodified forward primer.

All PCRs were carried out in 96 well plates in a total reaction volume of 10µl with reagent concentrations as outlined in Table 3.1. Initial optimization of PCR conditions involved running a 12 reaction temperature gradient running on the MJ research DYAD PCR machine where T_a ranged from 50 to 65°, such that a different annealing temperature was used for replicate samples. Each temperature gradient was carried out with final $MgCl_2$ concentration of 1.5mM and 2.0mM (Table 3.2).

For publicly available markers, the following cycling conditions were used: 95° for 5 minutes, 30 cycles of 95° for 30 seconds, T_a for 30 seconds and 72° for 30 seconds ending with a 5minute final extension at 72°. For some markers the number of cycles was increased to 34.

The novel markers MS18-1a, MS18-2, MS18-3, MS18-4a and MS18-5a were amplified in a multiplex PCR. The following cycling conditions were used: 95° for 15 minutes, 30 cycles of 95° for 30 seconds, 54.2° for one minute 30 seconds and 72° for 1 minute, followed by a final extension step of 60° for 30 minutes. Table 3.2 outlines PCR conditions, primer sequences and map location for all microsatellite loci used in this study.

Table 3.1 PCR components and concentrations for amplification of microsatellite loci.

Component	Concentration in reaction
PCR buffer	1x
$MgCl_2$[a]	1.5-2.5mM
dNTPs	0.2mM
Primer F[a]	0.15-0.45µM
Primer R[a]	0.15-0.45µM
Hot Fire Pol [b]	0.35U
DNA	~25ng
ddH20	to 10µl

a. See Table 3.2 for primer and $MgCl_2$ concentration information
b. From Solis Biodyne, Tartu, Estonia (http://www.sdb.ee)

Table 3.2 PCR conditions, primer pairs and map information for primer pairs included in this study.

Marker	cM (USDA)	Primer	5'-3' sequence	Dye	Ta (oC)	$MgCl_2$ conc. (mM)	Primer conc. (uM)
ILSTS002	54.713	F	TCTATACACATGTGCTGTGC	TAMRA	54.6	1.5	0.3
		R	CTTAGGGGTGAAGTGACACG				
DIK067	65.62	F	ACAGGTGATGCATGGGGTGG	TAMRA	56.4	1.5	0.3
		R	CCTCTTCCTAGTCACAGAGAC				
IDVGA-55	67.716	F	GTGACTGTATTTGTGAACACCTA	FAM	54.6	1.5	0.45
		R	TCTAAAACGGAGGCAGAGATG				
RME01	67.716	F	GTCTCCCTTCAGTGTTCCCA	FAM	56.4	1.5	0.225
		R	AGGTTGCTGGCTCCTTTTCT				
BMS2785	72.012	F	ACAAACCTGTGCGCCTTG	FAM	56.4	1.5	0.225
		R	GGCAATCAGTCGGACACAC				
UWCA5	74.521	F	CTTCCCTGTGTTAGGCTATGT	HEX	56.4	1.5	0.3
		R	CAAGAGAAGACAAGACTACAGA				
BM2078	76.782	F	CCCAAAAGAAGCCAGGAAG	FAM	58.9	2	0.45
		R	TCAGAGTTTGGGGTCCTCAG				
DIK5109	77.6	F	GCACAGAGATGCTCGGTCTC	HEX	58.9	2	0.3
		R	GCTCAGATCCTCCGCACTTA				
BM6507	78.845	F	ACTTAGCACAATGCCCTCTAGG	TAMRA	58.9	2	0.3
		R	ATGTTATTCCATCAGGAGGAGC				
MS18-5a	?	F	CTGAATGTAAAATAACCCTATGGG	HEX	54.2	2	0.3
		R	AACTAGAGAAAGCCAGTGCA				
MS18-3	?	F	TGGGGTCTGCTTTGAGGGCA	TAMRA	54.2	2	0.15
		R	GGAATTGCTGGATTGGGCA				
DIK5235	79.47	F	TCGTGTGCTGTTCATTGTCA	TAMRA	62.7	2	0.2
		R	TCAGCAGGCAGGTCCTTTAT				
KSCG	?	F	GTCACCTGTCTCAGTGCTTTG	FAM	58.9	2	0.21
		R	GGTCACCGGCTACAGGACAG				
TGLA227	84.087	F	CGAATTCCAAATCTGTTAATTTGCT	FAM	54.6	1.5	0.15
		R	ACAGACAGAAACTCAATGAAAGCA				
TGLA132	?	F	CCTTGGTTTCTAATTGTTTATTGCT	TAMRA	58.9	2.5	0.3
		R	ACATGACTTAGCAACTGAACAACAA				
RM128	73.53	F	GTGAGGAGGAGGGGCGTGGAG	HEX	58.9?	1.5	0.3
		R	AGGCCTGTGAGAAGATGGTGAG				
MS18-4a	?	F	CAAACCGCAAGGTCCTCATT	HEX	54.2	2	0.3
		R	CCCAGAGTCTGAGCAACCAA				

CentiMorgan positions refer to the USDA cattle linkage map (Ihara, 2004). The fluorescent labels used were FAM (blue), HEX (green) and TAMRA (yellow). $MgCl_2$ and primer concentrations listed are the concentrations in the final reaction volume. Forward and Reverse primers were used at the same concentration.

PCR products were electrophoresed on a 2% agarose gel stained with ethidium bromide at 0.5µg/ml and were visualized under UV light. Approximate concentration was determined and product size was confirmed by reference to 100bp ladder (Invitrogen). Publicly available microsatellites were amplified by PCR and run on 6% acrylamide gels on an ABI 377. Microsatellite data were collected using the GeneScan software and analysed using ABI PRISM GeneScan 2.1 (PE Applied BioSystems). Novel microsatellites were sent to the Genome Centre at St Bart's and the London, Queen Mary and Westfield College, London (http://www.smd.qmul.ac.uk/gc/), where electrophoresis was carried out on an ABI3700. Raw data were returned and was analysed using GeneMapper 4 (Applied Biosystems). Genotypic data were manually entered into an excel file. Microsatellite Toolkit (Park 2002) was used to check for errors, to calculate allele frequencies and basic diversity statistics, and to convert the data into GenePop and Microsat input formats.

3.2.4 Data analyses

3.2.4.1 Hardy Weinberg equilibrium

The Hardy-Weinberg Principle describes the relationship between allele frequencies and genotype frequencies in a population. A number of assumptions underlie this principle: the population is made up of random mating, diploid organisms; the allele frequencies are equal in males and females and reproduction is sexual, the population size is large and there is no introduction of new alleles by mutation or migration, and all alleles have equal fitness. When the assumptions above hold true, the allele frequencies and genotype frequencies do not change over time, even if the population size changes. The expected allele frequencies can be determined from genotype frequencies and vice versa. For genotype, A_iA_j, with a frequency X_{ij}, the frequency of the i^{th} allele is give by:

$$x_i = X_{ii} + \frac{1}{2}\sum_{j \neq i} X_{ij}$$

Given the allele frequencies for a locus with m alleles, the frequencies of each genotype can be determined by an expansion of:

$$(x_1 + x_2 + ... + x_m)^2$$

such that the genotype A_iA_j has a frequency of $2x_ix_j$ and the genotype x_{ii} has a frequency of x_i^2.

Deviations from Hardy-Weinberg Equilibrium (HWE) can be indicative of genotyping errors, null alleles, or violation of the assumptions on which the principle is based.

An exact test for Hardy Weinberg equilibrium (HWE) for each locus/population combination was carried out using a web based version of GenePop, available at http://genepop.curtin.edu.au/ (Raymond and Rousset 1995).

3.2.4.2 Gene diversity

Gene diversity was calculated using the Microsatellite Toolkit (Park 2002). This is Nei's unbiased gene diversity, and for diploid organisms is equivalent to expected heterozygosity (i.e., the proportion of heterozygotes expected in the population given the population size and the number of alleles at that locus, assuming HWE) and is a useful measure of diversity as it is not very sensitive to sample size (Nei 1987).

Gene diversity of a population can be estimated from a sample of that population by:

$$\hat{h} = 2n(1 - \sum \hat{x}_i^2)/(2n-1)$$

where n is the sample size, and $\hat{x}_i$ is the sample frequency of the i^{th} haplotype in the sample, given by:

$$\hat{x}_i = \hat{X}_{ii} + \sum_{j \neq i} \hat{X}_{ij}/2$$

where $\hat{X}_{ij}$ is the frequency of haplotype i,haplotype j in the sample.

3.2.4.3 Genetic distance

Nei's genetic distance and $(\delta\mu)^2$ were calculated for each locus between each population pair combination using Microsat. Nei's standard genetic distance is based on the infinite allele model of mutation, whilst $(\delta\mu)^2$ is based on the stepwise mutation model.

For two populations, *A* and *B*, Nei's standard genetic distance for any locus is given by:

$$d_S = -\ln\left(\frac{j_{AB}}{(j_A j_B)^{\frac{1}{2}}}\right)$$

where:

$$j_{AB} = \sum a_i b_i$$

$$j_A = \sum a_i^2$$

and

$$j_B = \sum b_i^2$$

a_i and b_i are the frequencies of allele the allele i in populations A and B. j_{AB} is the probability of an allele chosen at random from each of populations A and B being identical. j_A and j_A are the probabilities of two alleles drawn at random from the same population (A and B respectively) being identical.

For two populations A and B, $(\delta\mu)^2$ is given by:

$$(\delta\mu)^2 = (\mu_A - \mu_B)^2$$

μ_A and μ_B are the mean number of repeats for alleles in populations A and B (Goldstein et al. 1995).

Genetic distances were calculated using MICROSAT version 1.5 (Minch).

3.2.4.4 Observed and expected heterozygosity

Bottleneck (Cornuet and Luikart 1996) was used to calculate H_{OBS}/H_{EXP}, the difference between the observed heterozygosity and the expected heterozygosity under either infinite alleles model, the strict stepwise mutation model or the two phase mutation model, for each locus/population combination. This can be thought of as a reverse Ewens Watterson test, in which observed and expected Watterson's F are compared. F is equivalent to the expected homozygosity, and therefore the deviations from neutrality will have the opposite effect in a true Ewens-Watterson test compared to the test used here. A higher observed heterozygosity (and therefore lower observed homozygosity) than would be expected under neutrality, such as might occur if balancing selection were influencing a locus, would result in a increased value of H_{OBS}/H_{EXP} in the test used here (but a decreased value of F_{OBS}/F_{EXP} in the true Ewens-Watterson test). Conversely a lower observed heterozygosity (and higher observed homozygosity) as would be expected to occur if positive selection were influencing a given locus, would result in a decreased value of H_{OBS}/H_{EXP} (but an increased value of F_{OBS}/F_{EXP} in the true Ewens-Watterson test).

The observed heterozygosity is calculated directly from the number of alleles at the locus and number of samples genotyped, while the expected heterozygosity is an average value calculated from the distribution of gene diversity, given the number of alleles and the sample size.

Bottleneck was also used to calculate DH/sd, the difference between the observed heterozygosity and the expected heterozygosity, divided by the standard deviation of the expected heterozygosity.

3.2.5 Additional data

Additional microsatellite genotype data were available. The following loci had previously been genotyped in the same individuals from the same populations, with one exception mentioned below (Park 2002). However, not all loci were genotyped in all six populations. The loci were: ETH2, BMS597, BM733, IDVGA9, BMS8126, BMS772, TGLA377, CSSM042, BM3010, ILSTS030, BMS803, BL1001, BMS1300, RM356, BMS2024, CSSM005, BM1815, RM033, UWCA1, BM1258, BOLA1, DRB3, TAMLS113.3, BMS468, CYP21, MGTG7, RM185, BM7233, BM1818, BP34, CSSM024, BMS2269, BMS713, RM006, BM2607, BM1116, IL4, TGLA48, ILSTS001, BL1067, RM012, BM6105, TGLA303, HEL9, INRA005, INRA063, ETH10, ILSTS005, HEL1, B_ETH225, B_CSSM005, B_BMS2269, HBB, BM2113, RBP3, PRL, ETH131, His1, ILSTS014, INRA032, HEL5, OCAM, ETH152, RASA, TGLA116, CSSM066, INRA023, ILSTS006, CSRM60, BM1824, and Hel13 (Park, 2001).

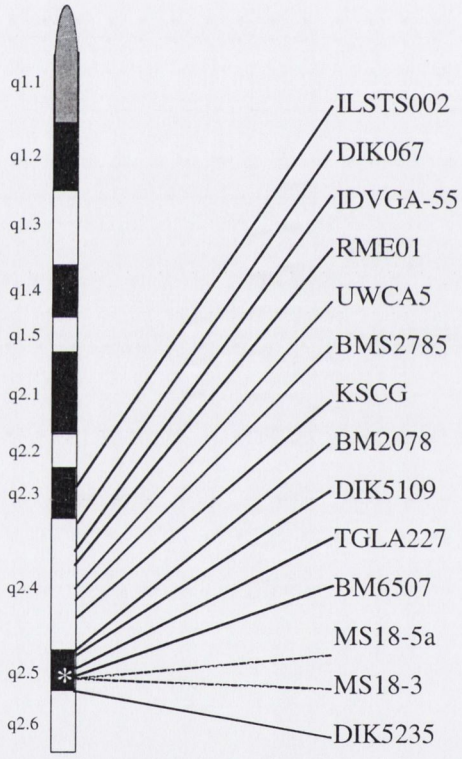
Genotype data for a further three loci genotyped in the same populations were also available (Freeman, AR, unpublished). These loci were DIK2027, DIK2989 and DIK4091.

3.3 Results

3.3.1 Data generated

Fourteen of the 17 microsatellites were successfully amplified and genotyped and these were included in the analysis. Of the three that failed, two RM128 and MS18-4 were excluded due to uncertainties with regards to allele calling. The exclusion of the third marker, TGLA132 is discussed in section 3.3.3. The microsatellite loci DIK5109 and DIK067 were not genotyped in the Somba population sample due to technical difficulties with PCR amplification of these loci. Figure 3.3 shows the approximate positions and relative order on BTA18 of the 14 successfully amplified microsatellites.

Figure 3.3 Schematic of microsatellites successfully amplified from BTA18q.



The microsatellite loci studied here are all dinucleotide repeats (Table 3.3). From inspection of GenBank/UniSTS entries, which were available for eleven of the loci, and bovine genome scaffolds for the novel markers, eleven loci appear to be made up of perfect dinucleotide repeats, while the remaining three loci contain an imperfect repeat, e.g. RME01 $(CA)_{10}CG(CA)_7$. No database information was available for the locus KSCG, but the alleles observed are consistent with it being a dinucleotide repeat. The alleles observed at the locus MS18-4a were inconsistent with the mutation method involving only the gain or loss of one or more complete dinucleotide, as expected under the SMM. Several alleles differed from the next closest in size by only 1bp.

Table 3.3 Microsatellite loci used in study.

Marker	GenBank	UniSTS	Repeat motif	Allele size range
ILSTS002	L23479	250878	$(AC)_{17}$	121-137
DIK067	D83290/D83290	251566	$(TG)_7TA(TG)_7$	91-121
IDVGA-55	X85071	251462	$(TG)_{12}$	192-204
RME01	U15431	251472	$(CA)_{10}CG(CA)_7$	156-172
BMS2785	G18989	6954	$(TG)_{15}$	103-121
UWCA5	L15334	250758	$(CA)_8$	123-125
BM2078	G18411	31043	$(AC)_{15}$	92-118
DIK5109	AB166161	N/A	$(TG)_{11}$	192-204
BM6507	G18441	10740	$(AC)_{16}$	147-165
MS18-5a	SCAFFOLD291578		$(AC)_{13}$	216-240
MS18-3	SCAFFOLD10653		$(AT)_6GCAC(AT)_{13}$	238-260
DIK5235	AB166257	N/A	$(AC)_{16}$	185-209
KSCG	N/A	N/A	Unknown	262-288
TGLA227	N/A	250914	$(AC)_{19}$	80-106

Marker name, GenBank Accession ID and UniSTS ID where available, repeat motif and observed allele size range for the fourteen microsatellites successfully genotyped.

3.3.2 Genomic positions of the microsatellites.

3.3.2.1 Genomic positions of the publicly available microsatellites

Genomic positions of the microsatellites were determined using *in silico* PCR (http://genome.ucsc.edu) on the August 2006 assembly of the bovine genome (Btau_3.1) (Table 3.4). Positions in bp were converted to Mb and rounded to two decimal places for the purpose of subsequent analyses. The position of ILSTS002 was determined by a BLAT search (http://genome.ucsc.edu) using the GenBank sequence X85071.

Table 3.4 Genomic positions of microsatellites used in study.

Marker	start (bp)	end (bp)	Mb position
ILSTS002	40818951	40819037	40.82
DIK067	49188844	49188936	49.19
IDVGA-55	51286393	51287013	51.29
RME01	51723128	51723297	51.72
UWCA5	53373296	53373418	53.37
BMS2785	53952952	53953062	53.95
KSCG	54648491	54648757	54.65
BM2078	59003295	59003402	59
DIK5109	59469261	59496458	59.47
TGLA227	60360145	60360234	60.36
BM6507	61221291	61221451	61.22
MS18-5a*	NA	NA	61.3
MS18-3*	NA	NA	61.35
DIK5235	61592237	61592424	61.6

Microsatellite positions as determined by reference to the current release of the bovine genome (Btau_3.1). Start and End bp positions given for each marker indicate the start and end of the PCR product. A position in Mb is also given. NA indicates that the microsatellite could not be mapped to the genome assembly. Novel microsatellite loci are indicated by *.

3.3.2.2 Genomic positions of novel microsatellites

It was not possible to determine the exact positions of the novel microsatellites using the *in silico* PCR approach due to gaps in the genomic sequence in this region of the chromosome.

The order of the novel microsatellites, relative to each other was estimated based on available information. MS18-3 lies on the same genomic scaffold as the FcaR gene, which in human, chimpanzee and rat, lies at the telomeric end of the LRC. MS18-4a is located close to the KIR2DL1 gene, which is the next gene centromeric to the FcaR gene. Therefore, the marker order, from centromeric end to telomeric end is estimated to be MS18-5a, MS18-4a, MS18-3.

The position of the novel microsatellite loci, relative to the publicly available microsatellite loci was unknown and was estimated by blasting GenBank nucleotide sequences for the microsatellite loci BM6507 and DIK5235 against the bovine genome sequences at the ensembl website and looking at regions of homology between the bovine and human genomes. The BM6507 sequence lay close to a region of homology in which the gene LAIR2 is located. This gene is part of the LRC and is centromeric to the KIR gene cluster. The DIK5235 sequence lay within a region of homology in which no human genes are located. This region ranged from 61,330,181bp to 61,335,180 on HSA19, telomeric to the KIR genes. The human FcaR gene was located at 60,095,055bp.

In the absence of any further information the microsatellite loci MS18-5a, MS18-4a and MS18-3 have been assigned to the region between BM6507 and DIK5235. As gaps in the genome sequence are filled, it should become possible to refine these positions. For now, these microsatellites have been assigned arbitrary positions (*) between BM6507 and DIK5235.

3.3.3 HWE

The test of HWE as implemented by GenePop, was run for every possible population/locus combination. This test identified one microsatellite locus, TGLA132, which deviated significantly at the level $p<0.005$ from expectation assuming HWE in three of the four populations tested. This marker was monomorphic in the two European *Bos taurus* populations and therefore no test was carried out in these populations. In the Nellore, N'Dama and Somba populations there was an excess of homozygotes as indicated by a positive F_{IS} value, however it showed no deviation from HWE in the Ongole population. Further investigation revealed that every male sample typed for this marker was homozygous. The genome scaffold on which this marker lay was examined in more detail, and the nearest gene identified as PHD finger protein 8, which has been mapped to the X chromosome. Unlike the other populations, the Ongole population sample was made up primarily of female samples, and included only three male samples in a total of thirty. For this reason the test of HWE showed no deviation from the expectation under neutral evolution in the Ongole sample population.

Of the remaining microsatellite loci, five locus/population comparisons showed significant deviation from expectations assuming HWE at the $p<0.05$ level (Table 3.5) None of these were significant at the $p<0.01$ level or when a Bonferroni correction for multiple testing was applied.

Table 3.5 Deviations from Hardy-Weinberg Equilibrium

Population	Marker	p-value
Charolais	MS18-3	0.0388
Friesian	DIK067	0.0109
Ongole	KSCG	0.0217
Somba	IDVGA-55	0.015
Somba	BMS2785	0.027
Somba	TGLA132	<0.0001
N'Dama	TGLA132	0.0019
Nelore	TGLA132	<0.0001

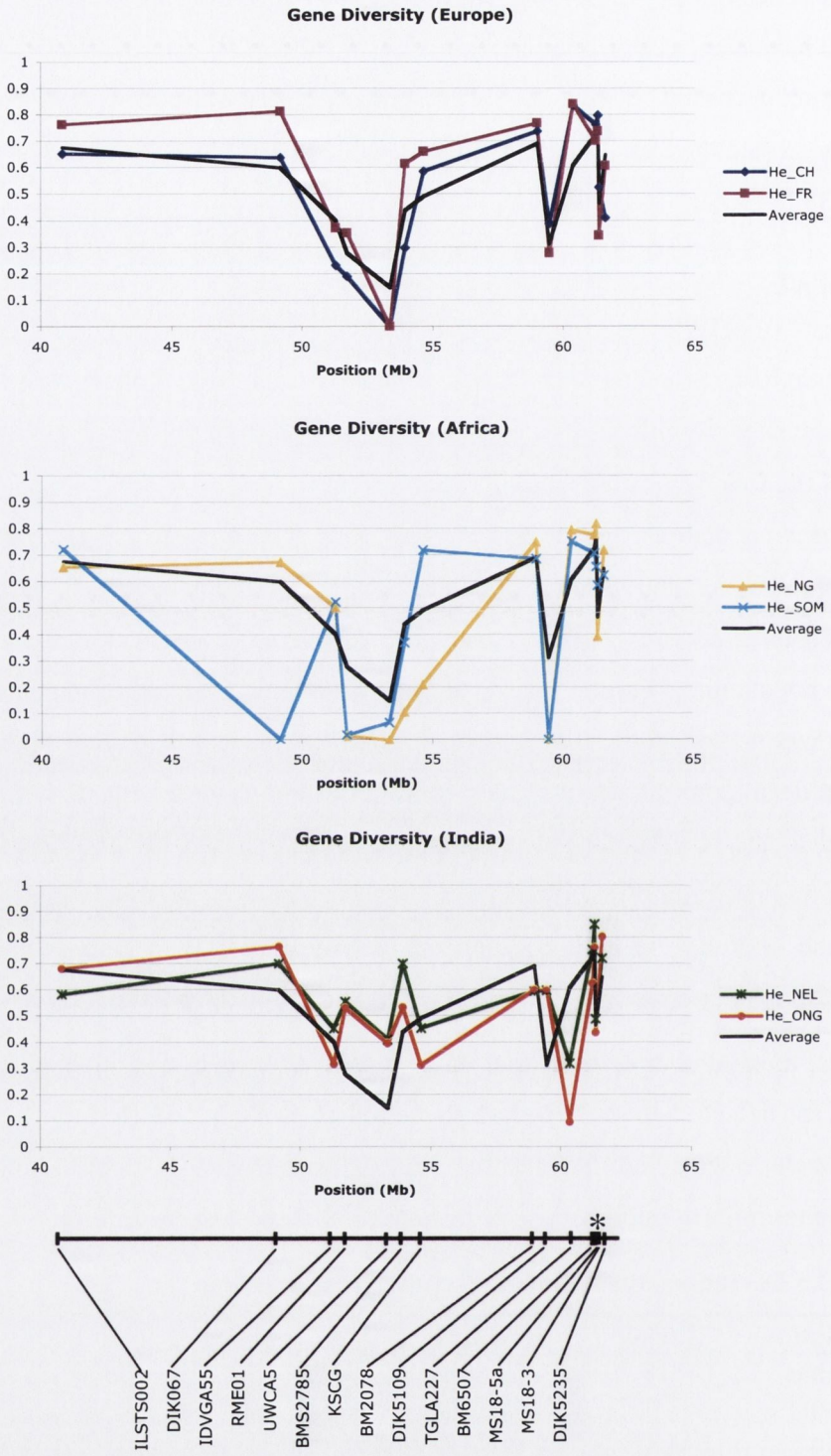


Figure 3.4 Gene diversity for 14 loci on BTA18.

Gene diversity (heterozygosity) plotted against chromosomal position for two European *Bos taurus* breeds (top), two African *Bos taurus* breeds (centre) and two *Bos indicus* breeds (bottom). The average gene diversity across all breeds at each locus is shown as a black line. The approximate position of the KIR genes is indicated by *.

3.3.4 Gene Diversity

Gene diversity is equivalent to expected heterozygosity in the case of diploid organisms and is a useful measure of diversity as it is less sensitive to sample size than other measures such as number of alleles (Park 2002). In previous work with the same population sample, it has been shown that gene diversity for microsatellite loci along BTA23 peaks around the MHC, consistent with balancing selection acting on these loci. Gene Diversity was calculated using the Microsatellite Toolkit. Figure 3.4 illustrates the average intra-continental gene diversity.

Although there is no obvious trend emerging from these data, gene diversity appears to be greatest in the *Bos indicus* breeds, the Nellore and the Ongole, lowest in the African *Bos taurus* breeds and intermediate in the European *Bos taurus* breeds.

The locus with the lowest average diversity is UWCA5. This is primarily due to the microsatellite being monomorphic in the Charolais, Friesian and N'Dama. Two alleles were detected in the Nellore, Ongole and Somba. The minor allele frequency in the Somba was 3.33%. This allele is absent from all other *Bos taurus* breeds studied here, and is present at over seventy percent in both of the *Bos indicus* breeds. This observation is not unexpected, as it has been previously shown that the *Bos taurus* Somba population has experienced minor *Bos indicus* introgression (Freeman, 2003).

The loci with the highest average diversity are the neighbouring loci BM6507 and MS18-5a. All six populations exhibit high diversity at these loci. MS18-5a lies approximately 8kb from KIR2DS1 gene. The distance between MS18-5a and BM6507 is unknown.

3.3.5 Genetic Distances

Nei's genetic distance and $(\delta\mu)^2$ squared were calculated for each pairwise population comparison at each locus using Microsat. Nei's genetic distance is based on the infinite allele model, while $(\delta\mu)^2$ is based on the stepwise mutation model. Figure 3.5 illustrates the average genetic distance over i) all *Bos taurus-Bos indicus* pairwise distances, ii) all European *Bos taurus* – African *Bos taurus* distances and iii) all intra-continental distances.

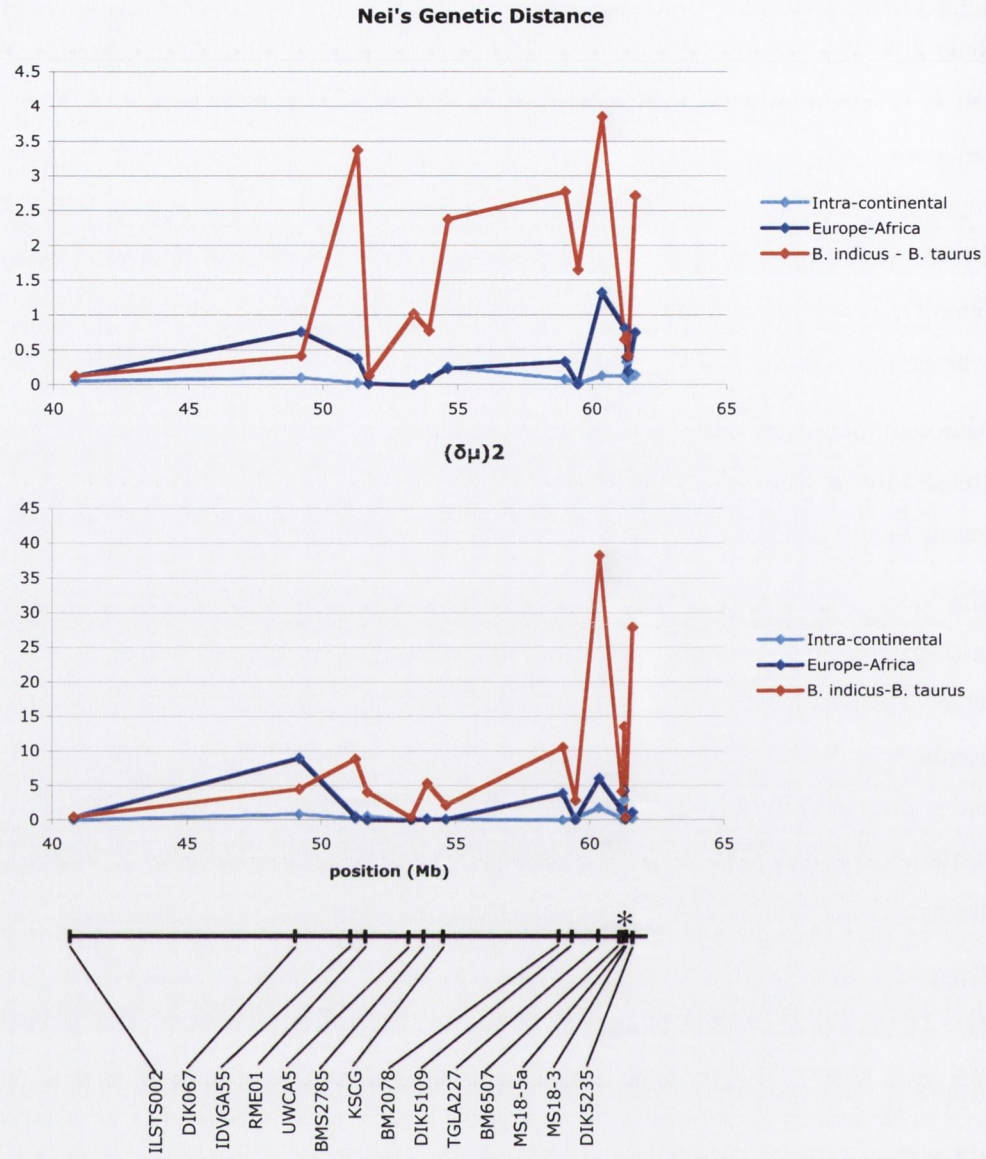


Figure 3.5 Genetics distances between populations at 14 loci on BTA18.

Nei's standard genetic distance (top) and $(\delta\mu)^2$ (bottom) are given as averages of all pairwise distances for each category (3 pairwise comparisons within continents, 4 pairwise comparisons between African *Bos taurus* and European *Bos taurus* and 8 pairwise comparisons between *Bos indicus* and *Bos taurus*. The approximate position of the KIR cluster is indicated by *.

Nei's standard genetic distance (top) and $(\delta\mu)^2$ (bottom) between populations at each locus included in the study are shown in Figure 3.5. Distances illustrated are the average of all pairwise distances for each category; 3 pairwise comparisons for the intra-continental distance (blue), 4 pairwise comparisons for the African *Bos taurus* – European *Bos taurus* distance (purple) and 8 comparisons for the *Bos taurus* – *Bos indicus* distance (red). The position of each locus is shown by the schematic below the chart, with the approximate location of the KIR cluster indicated by an asterisk.

The intra-continental distance, the average distance between breeds originating on the same continent, is the smallest, the *Bos indicus-Bos taurus* distance is greatest, and the distance between African and European *Bos taurus* breeds is intermediate. This is consistent with what is known about the origins of domestic cattle. There are however some loci that do not show the same pattern. At the locus DIK067, the *Bos indicus-Bos taurus* distance is less than the African-European *Bos taurus* distance.

Overall, a similar pattern is observed for the two distance measures, however there are some notable differences in the *Bos indicus-Bos taurus* distances between the two graphs. The highest peaks in the *Bos indicus-Bos taurus* distance in the graph of Nei's standard genetic distance (Figure 3.5) are likely to be due to the fact that the most frequent allele in *Bos indicus* is at very low frequency or absent in the *Bos taurus* breeds. Two of these peaks, at IDVGA-55 and BM2078, are much decreased when the $(\delta\mu)^2$ distance (based on the stepwise mutation model) is considered, while the other two, DIK5235 and TGLA227, remain high. This apparent discrepancy may due to the influence of differences in the allele size distribution on the distance measure. The $(\delta\mu)^2$ distance measure is based on the SMM model and is largely concerned with the mean allele size, whereas Nei's standard genetic distance is based on the IAM and is dependent on allele identity between populations. Under a SMM model, the number of mutations separating two alleles that differ by multiple units of the repeated motif is typically greater than the number of mutations separating the same two alleles under the IAM. This is because according to the SMM, each mutation results in the gain or loss of a single unit of the repeated motif. Figure 3.6 illustrates the differences in the allele frequency distributions for DIK5325 (left) and BM2078 (right) in European *Bos taurus* (top), African *Bos taurus* (middle) and European *Bos taurus* (bottom). The most frequent alleles in the *Bos indicus* populations, indicated by red arrows, are absent or present at low frequency in the *Bos taurus* populations. At the microsatellite DIK5235, these alleles are separated by a number of repeat units from the more frequent alleles in the *Bos taurus* populations In

contrast, At the locus DM2078, the neighbouring allele (96) is present at a relatively high frequency in Aftican *Bos taurus* and at an appreciable frequency in relative to the other loci European *Bos taurus*. The $(\delta\mu)^2$ distance measure, which remains high for *Bos taurus Bos indicus* comparisons in DIK5235 but is greatly reduced for BM2078, illustrates this.

Figure 3.6 Allele frequency distributions for DIK5325 (left) and BM2078 (right).

The allele sizes shown are the length of the PCR product. The alleles with the highest frequency in *Bos indicus* is highlighted with red arrows. These alleles are at low frequency or absent from the *Bos taurus* populations.

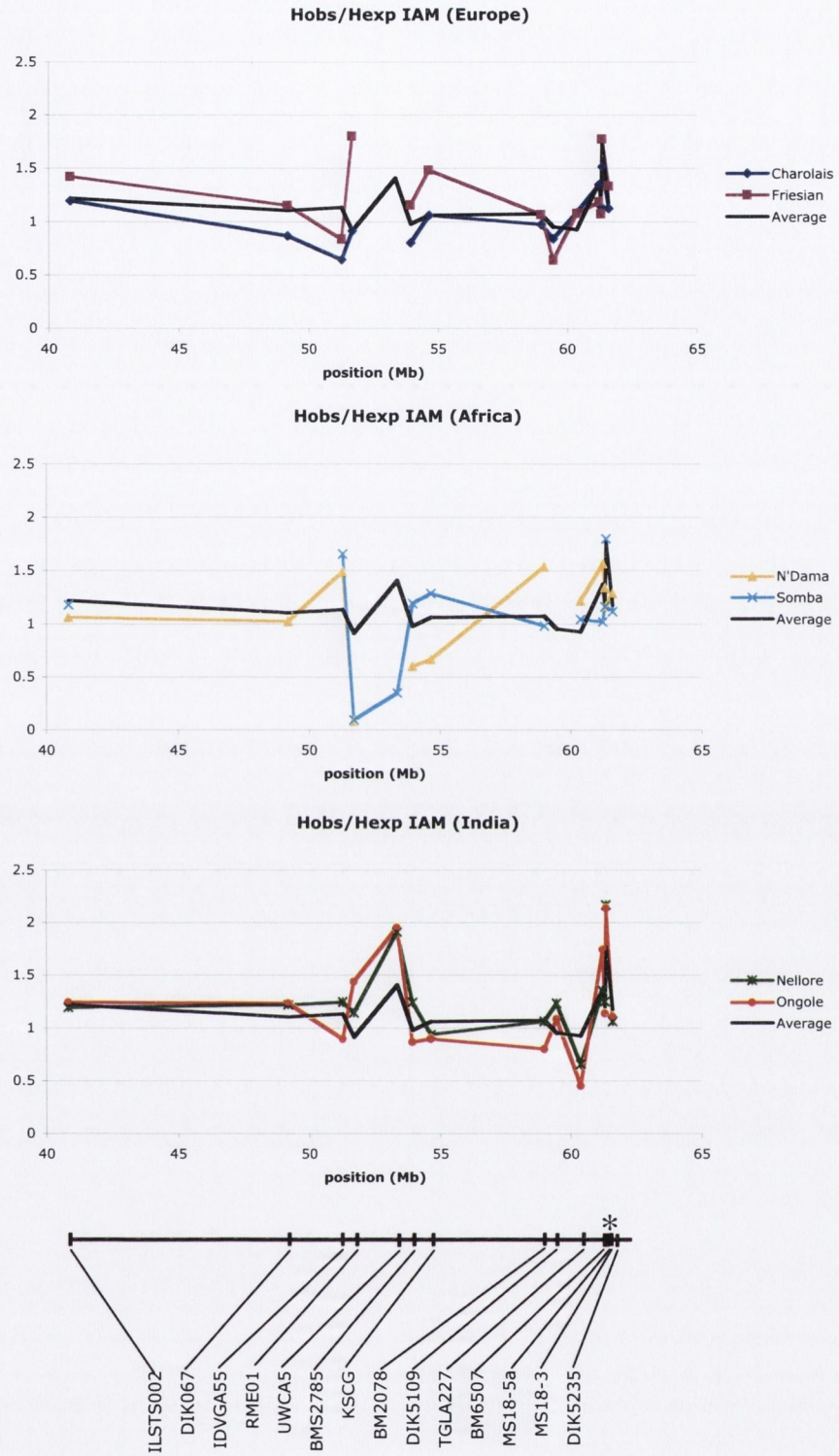


Figure 3.7 Ratio between the observed and expected heterozygosity.

H_{OBS}/H_{EXP} plotted against chromosomal position for two European *Bos taurus* breeds (top), two African *Bos taurus* breeds (centre) and two *Bos indicus* breeds (bottom). The average value of H_{OBS}/H_{EXP} across all breeds at each locus is shown as a black line. The approximate position of the KIR genes is indicated by *.

3.3.6 Ewens Watterson based tests (H_{OBS}/H_{EXP} and DH/sd)

Bottleneck estimates the expected heterozygosity assuming one of three mutation models, the IAM, the SMM, or the TPM. The data from this study was run under each model, and although the absolute values for the expected heterozygosity were not the same, the data from each run showed the same trend. Figure 3.7 shows the H_{OBS}/H_{EXP}; the ratio between the observed and expected heterozygosity under the IAM for a) the European breeds, b) the African breeds and c) the Indian breeds. The black line one each graph shows the average across all breeds. Missing points exist where a marker is monomorphic in that breed.

Positive selection, which would increase the frequency of the selected allele, would result in an decreased observed heterozygosity relative to the expectation under neutral evolution, and therefore a lower value for H_{OBS}/H_{EXP} whereas balancing selection, which would give rise to a more even allele frequency distribution, would result in a higher value of observed heterozygosity, and therefore a higher H_{OBS}/H_{EXP} ratio. Under neutral evolution, the value of H_{OBS}/H_{EXP} is expected to be 1.

Once again, no clear pattern of the variation between observed and expected heterozygosity along this region of BTA18 emerges from this study. There is considerable variation in H_{OBS}/H_{EXP} between different breeds at the same locus. On average (black line) the highest values of Hobs/Hexp are found at the loci closest to the KIR region, but this observation is not striking against the pattern of variation along the rest of the chromosome.

The measure DH/sd was also calculated for each locus in each population (Figure 3.8). DH/sd is the difference between the observed and expected heterozygosity, divided by the standard deviation of the expected heterozygosity. Under neutral evolution, the value of DH/sd is expected to be close to zero. Because the standard deviation of the expected heterozygosity decreases as the number of alleles per locus increase, DH/sd is expected to be less sensitive to variation in number of alleles per locus.

DH/sd shows great variation, both between microsatellite loci in the same population and between populations at the same locus. The highest average DH/sd was found at BM6507 (1.16), MS18-3 (1.10) and MS18-5a (0.988).

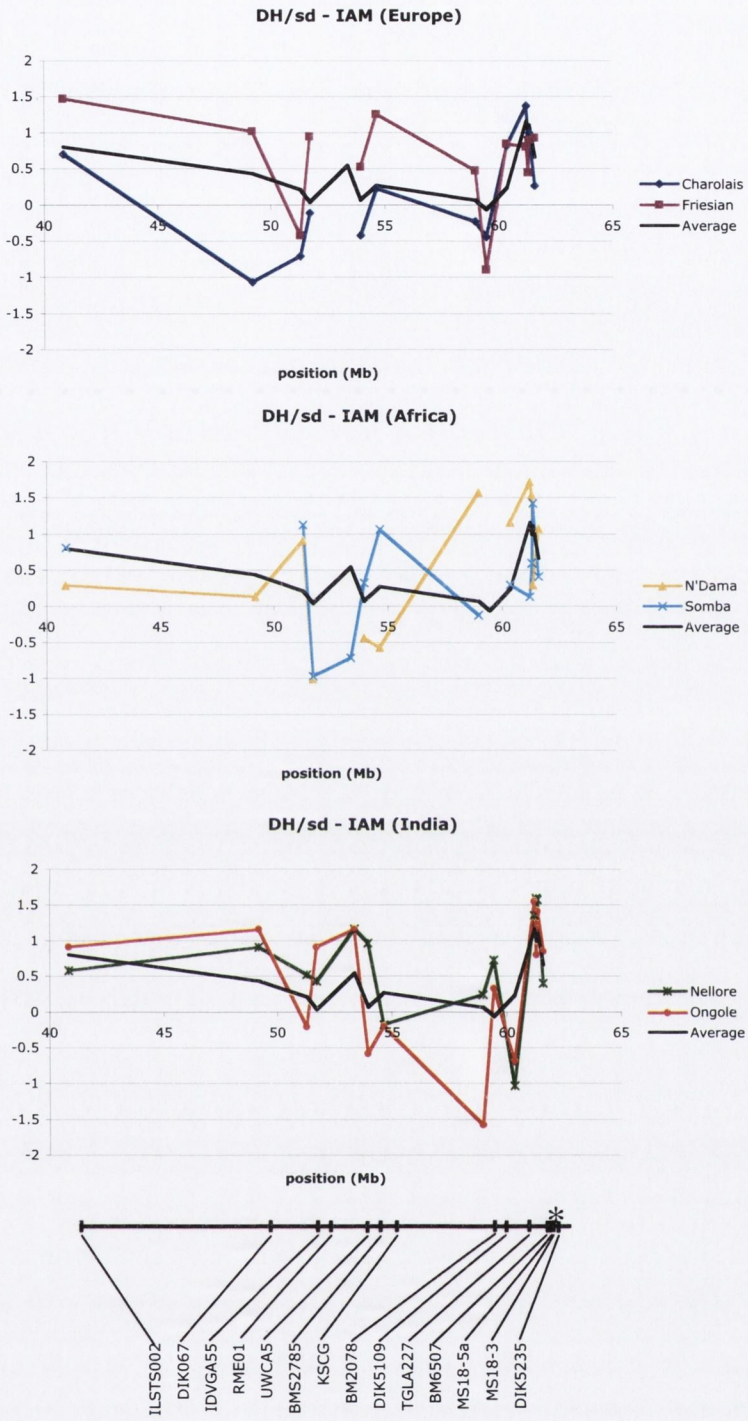


Figure 3.8. DH/sd plotted against chromosomal position.

DH/sd plotted against chromosomal position for two European *Bos taurus* breeds (top), two African *Bos taurus* breeds (centre) and two *Bos indicus* breeds (bottom). DH/sd is the difference between the observed and expected heterozygosity, divided by the standard deviation of the expected heterozygosity. The average value of DH/sd across all breeds at each locus is shown as a black line. The approximate position of the KIR genes is indicated by *.

3.4 Discussion

3.4.1 Comparison of BTA18 microsatellite data with similar data surrounding the MHC.

A comprehensive investigation of microsatellite variation at and around the MHC was carried out by Park (2002). Five MHC-linked loci along with 13 loci spanning almost 55cM on the same chromosome, BTA23, were genotyped in cattle from six population representing European *Bos taurus*, African *Bos taurus* and Indian *Bos indicus* animals. In addition, genotype information for a further 49 microsatellite loci spread around the genome was also used, allowing comparisons to be made between MHC-linked loci and flanking loci and between MHC-loci and a larger number of unlinked loci around the genome. In total, 67 microsatellites were examined, and these were classified as MHC-linked (n=5), gene linked (n=8) or anonymous (n=54).

Nei's standard genetic distance was calculated for each pairwise combination of populations at each locus. The greatest distances were observed between *Bos indicus* and *Bos taurus* populations, and the shortest distances between populations from the same continents. The distances between European and African *Bos taurus* animals were intermediate (Figure 3.9, centre). For MHC-linked loci, the *Bos indicus-Bos taurus* distance and the African-European distances were reduced compared with flanking loci, while the intra-continental distances were elevated suggesting that the MHC-linked loci have been affected by balancing selection. The average gene diversity, which is equal to the expected heterozygosity, across all populations was significantly greater for MHC-linked loci compared to anonymous or gene-linked loci (Figure 3.9, bottom). The allele frequency distributions at each locus were examined, using the Ewens Watterson test. The Ewens Watterson's test is based on Ewens sampling formula and Watterson's F statistic, a measure of gene diversity equivalent to homozygosity. Park and colleagues examined the variation in F_{OBS}/F_{EXP} across a large region encompassing the bovine MHC. Deviations from the expected F give some evidence that a locus may be under selection. Balancing selection results in a more even distribution of allele frequencies and therefore a lower value of F_{OBS}, while directional selection has the opposite effect. MHC-linked loci showed more even allele frequencies and therefore a lower F_{OBS}/F_{EXP} than anonymous loci (Figure 3.9 top). Since the F statistic is equivalent to homozygosity, the opposite effect would be expected compared to H_{OBS}/H_{EXP}, that is, an increased value where positive selection has been acting, and a decreased value where balancing selection

has been acting. A sharp dip is seen at MHC-linked loci and the difference between F_{OBS}/F_{EXP} at MHC-linked loci and non-MHC-linked loci studied was highly statistically significant (Park 2002).

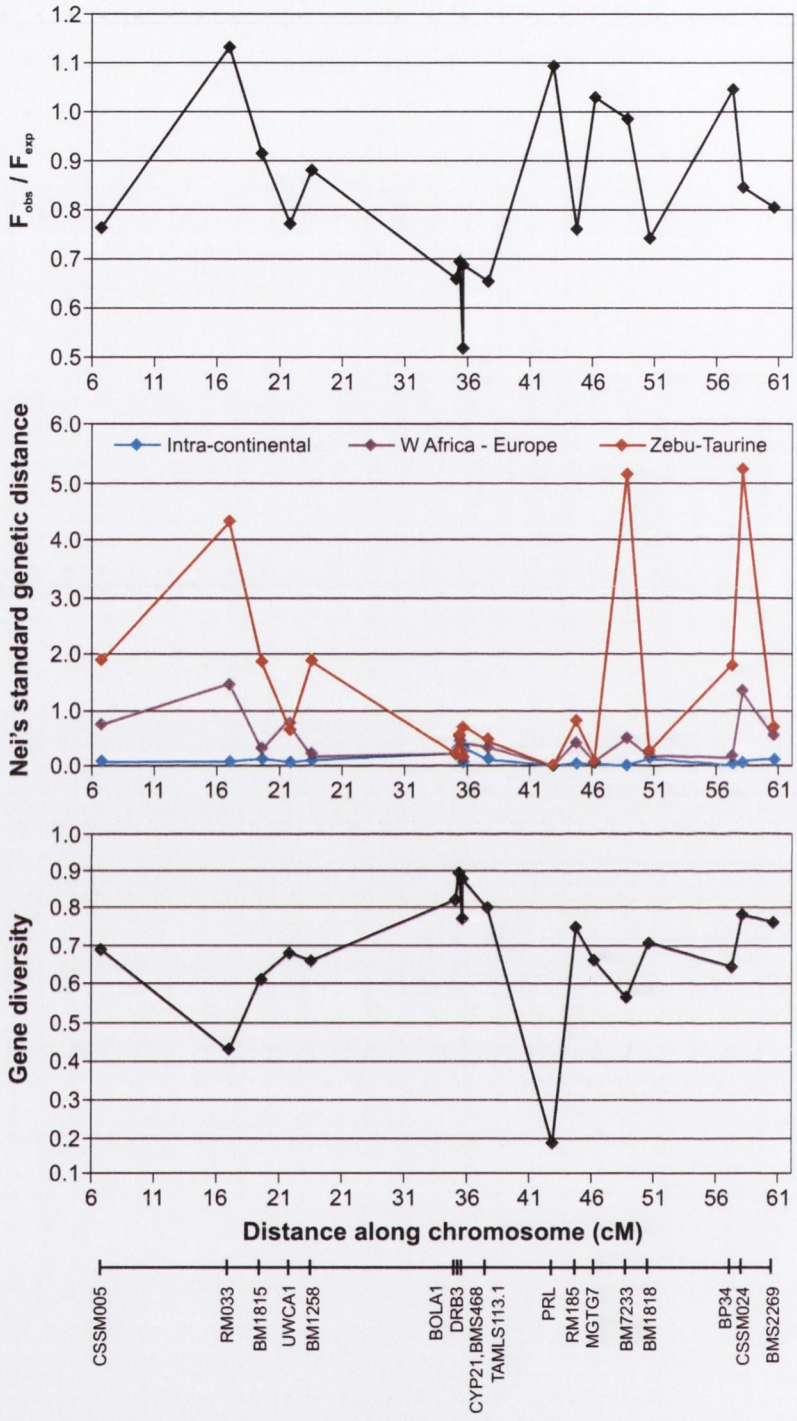


Figure 3.9 F_{OBS}/F_{EXP}, genetic distance and gene diversity at 18 loci on a region BTA23 encompassing the MHC.

Fobs/Fexp and gene diversity values are averages across all populations and genetic distance values are averages of all pairwise comparisons in each class. The MHC-linked microsatellites are BOLA1, DRB3, CYP21, BMS468 and TAMLS113.3. (Figure redrawn from Park, 2002).

Park and colleagues examined the variation in F_{OBS}/F_{EXP} across a ~60cM region encompassing the bovine MHC. This F statistic is equivalent to homozygosity, and therefore the opposite effect would be expected compared to H_{OBS}/H_{EXP}, that is, an increased value where positive selection has been acting, and a decreased value where balancing selection has been acting. A sharp dip is seen at MHC-linked loci and the difference between F_{OBS}/F_{EXP} at MHC-linked loci and non-MHC-linked loci studied was highly statistically significant (Park 2002).

3.4.2 Comparison with a genome-wide microsatellite dataset

By considering the data from the current study together with the microsatellite data from Park (Park 2002) and additional genotype data from three microsatellites types in the same populations (Freeman, A.R., unpublished), it is possible to consider the data against an empirical distribution. In total, this combined dataset contains 87 microsatellite markers genotyped in the same six cattle populations. However, not all of these microsatellites have been typed in every population, with number of markers typed per population ranging from 48 for the Somba to 87 for the Charolais

This large dataset allowed a genomewide distribution of DH/sd to be constructed. This may be a more appropriate measure to use as dividing by the standard deviation takes into account the variation in number of alleles per locus. Standard deviation of expected heterozygosity decreases as the number of alleles per locus increases, therefore a deviation of the same magnitude is more significant at a highly polymorphic locus than at a less polymorphic locus.

Figure 3.10 shows the distribution of DH/sd for each of the six breeds included in the study. The positions of the 5 MHC-linked loci (fromPark 2002) in the distribution have been indicated by black arrows and the position of the 3 KIR-linked loci have been indicated by red arrows. MS18-3 and MS18-5a are known to be closely linked to genes in the KIR cluster as the tandem repeat sequence was identified on the same scaffold as an LRC gene, and BM6507 has been included on the basis of its position which was determined to be close to the KIR cluster and lying within the LRC.

The values of DH/sd for the KIR-linked loci, MS18-3, MS18-5a and BM6507 are comparable to those obtained for the MHC-linked microsatellites. For three of the six populations, an MHC-linked microsatellite has a higher value of DH/sd than all the KIR-linked microsatellites, but for the remaining three populations, a KIR-linked microsatellite has higher value than the MHC populations. However, in all populations,

the most extreme marker at the positive end of the distribution is neither MHC-linked, nor KIR-linked, nor is same marker consistently the most extreme marker.

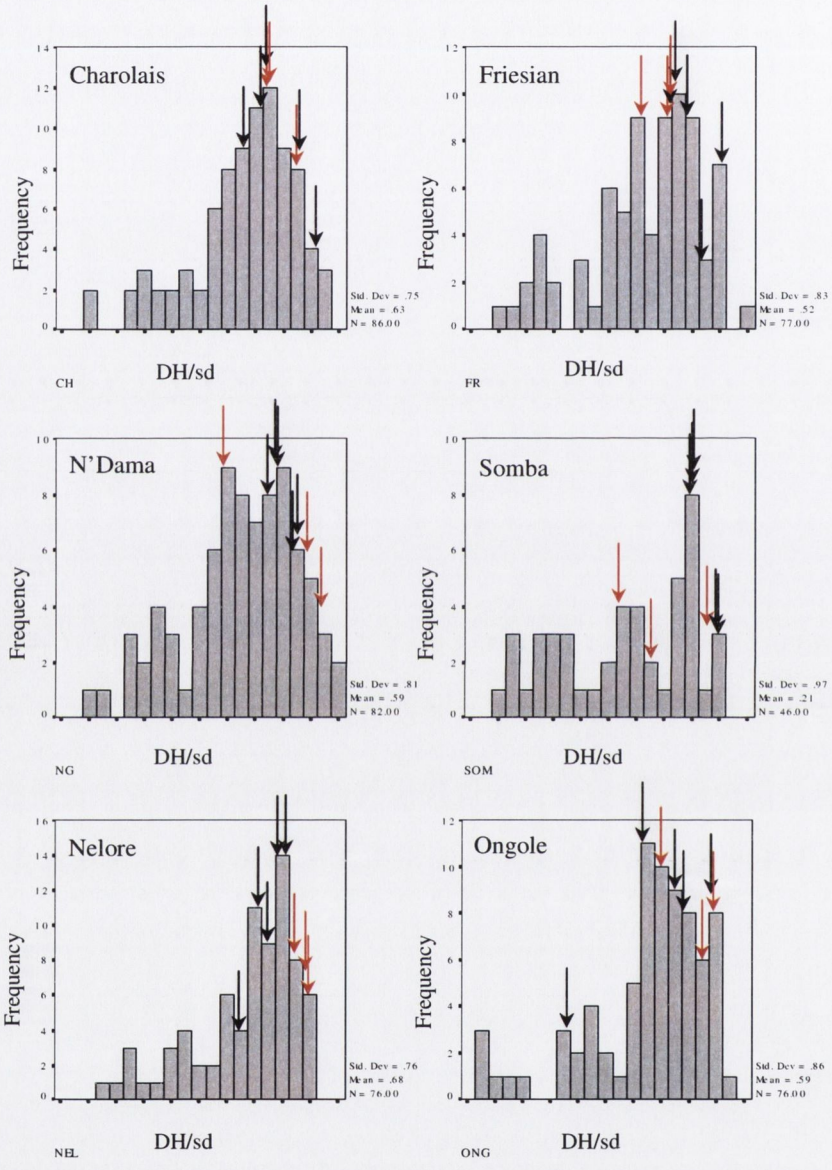


Figure 3.10 Distribution of DH/sd for each of the six breeds included in the study. The positions of the 5 MHC-linked loci (fromPark 2002) in the distribution have been indicated by black arrows and the position of the 3 KIR-linked loci have been indicated by red arrows.

By combining the values for all six populations such that the combined dataset has (up to) six values per microsatellite locus, the data can be viewed as three separate distributions. Figure 3.11 shows the distribution of DH/sd values for all non-KIR, non-MHC-linked loci

(top), all KIR-linked loci (middle) and all MHC-linked loci (bottom). This clearly illustrates that the DH/sd values for the MHC-linked loci fall at the positive end of the distribution and that the values for the loci for the KIR-linked loci, while being comparably extreme at the positive end of the distribution, show greater variance, approaching that of the non-KIR, non MHC distribution

The significance of the difference in dH/SD between the KIR-linked loci and non KIR, nonMHC loci and the MHC-linked loci and non KIR, non MHC was tested by Mann Whitney U test at the continental level and at the regional level. While the difference in dH/SD values between the MHC and non-KIR, non-MHC loci was significant for the Somba, N'Dama and Friesian populations, and was significant for all groupings at the continental level, no evidence for a significant difference between KIR-linked loci and non-KIR, non-MHC-linked loci was observed (Table 3.6).

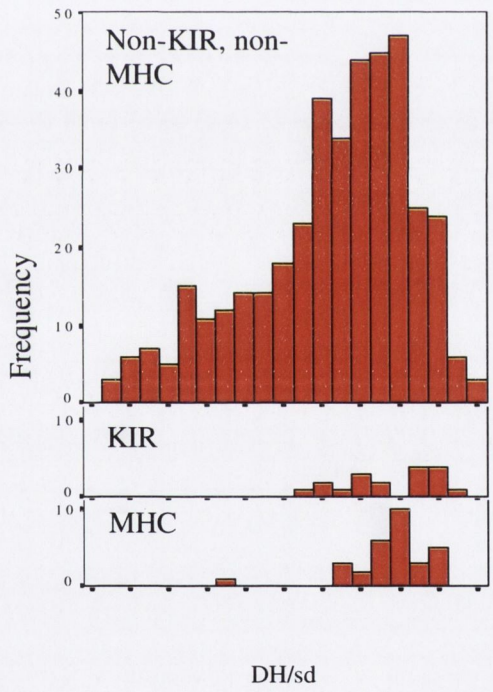


Figure 3.11 DH/sd distributions across all populations subdivided into non-KIR/MHC (top), KIR-only (middle) and MHC-only (bottom)

Table 3.6 Mann-Whitney U-Test of different DH/sd distributions in the six breeds.

Breed	KIR v rest	MHC v rest	KIR v MHC
Charolais	n.s	n.s	n.s
Friesian	n.s	0.036	0.036
N'Dama	n.s	0.032	n.s
Somba	n.s	<0.001	n.s
Nelore	0.002	n.s	0.036
Ongole	n.s	n.s	n.s

3.5 Conclusions

This analysis provides evidence of balancing selection acting on the KIR-linked microsatellites. The results are similar to, though less pronounced than, the evidence previously found for the MHC region (Park 2002), which has long been a classic example of proposed balancing selection in humans. The KIR region showed the highest average gene diversity, the highest average H_{OBS}/H_{EXP}, and the highest average DH/sd. Genetic distances between populations were also found to be low in this region. These observations are all suggestive of balancing selection. The signal of balancing selection at the KIR-linked loci is strongest in the *Bos indicus* populations despite a limited sample size. The difference between the DH/sd values of the KIR-linked microsatellites and the non-KIR, non-MHC-linked microsatellites was the only significant result observed. Balancing selection at the KIR cluster has already been reported in humans (Norman et al. 2004). It would not be surprising if cattle, which share the same major subdivisions of KIR genes as humans (Figure 3.2), are subject to similar selective forces. This is especially true in the time since domestication, when humans and cattle would have been subject to many of the same disease challenges. Furthermore, recent work has uncovered evidence for coevolution of the KIR and HLA genes in human populations (Single et al. 2007). This is entirely consistent with the close functional relationship between the KIR and MHC loci; indeed, it had previously been suggested that KIR genes may be subject to similar evolutionary forces as MHC loci (Martin et al 2002, Vilches and Parham, 2002). Although not conclusive, this work supports the hypothesis that the same forces are also at work in cattle.

Empirical genomic scans, such as presented here, have a number of advantages over pure comparisons to theoretical predictions but have also come under criticism for returning outlying loci that are unrepresentative of adaptive loci in general (Kelley et al. 2006; Teshima et al. 2006). The Ewens Watterson approach applied here to microsatellite data seems to be of some value, as it did identify signs of balancing selection in the KIR locus, which makes a lot of biological sense and fits well with existing data. Nevertheless, there are also signs that microsatellites are of limited use in detecting selection. Whereas groups of markers (i.e. KIR-linked, or MHC-linked) do tend to lie at the positive end of the distribution, differences between the DH/sd values of the KIR-linked and MHC-linked microsatellites and the non-KIR, non-MHC-linked microsatellites are only significant in some of the populations (3/6 for MHC and only 1/6 for KIR). The power to detect this effect at individual loci is smaller still. If the KIR genes had not been

previously identified as candidate loci for deviations from normality, it is not clear that they would have been identified from a genomic scan. For loci under weaker selection, the situation is likely to be bleaker still. This reinforces the need for caution when interpreting the extreme loci of any empirical genomic scans for selection, including those presented in Chapter 2 (Table 2.4).

4 Investigation of Candidate Genes by Sequence Analysis

4.1 Introduction

4.1.1 The Genes

Genes involved in the immune response have been found to be over-represented in a number of studies that aimed to detect evidence of selection (Nielsen et al. 2007). Given the obvious selective advantage of better resistance to disease challenges, this is not surprising. Domesticated species are thought to be even more likely to have undergone recent selection at these loci due to the epidemiological changes associated with the domestication process and subsequent migration into new habitats. Given the wide range of different disease challenges cattle must have experienced, it would not be surprising to find selection acting at a variety of immune response genes. Such genes therefore make excellent targets for the candidate gene approach

4.1.1.1 Toll like receptors

Toll like receptors (TLRs) are an evolutionarily conserved class of cell surface receptors whose ligands are pathogen associated molecular patterns (PAMPs) (Roach et al. 2005). Upon binding of the ligand, the TLR transduces the signal via the recruitment of adaptor molecules MyD88, MAL, TRIF and TRAM (O'Neill and Bowie 2007). Signalling through TLRs stimulates the innate immune response as well as activation of co-stimulatory molecules, which initiate the adaptive immune response. Given this direct interaction with PAMPs, and their position at the interface of the adaptive and innate immune responses, the TLR are prime candidates to have been affected by selection as a result of disease challenge. Indeed, variation in TLRs has been linked to increased susceptibility or resistance to infection (White et al. 2003). Expression of TLR2 and TLR4 was found to be increased in mastitis infected udders (Goldammer et al. 2004). Mastitis is one of the most economically damaging diseases in dairy cattle.

Toll-like receptor 2 (TLR2) recognises a wide variety of ligands originating from gram positive bacteria, including peptidoglycan, lipoproteins, lipoteichoic acid. TLR2 forms heterodimers with TLR1 and TLR6 which may explain the wide range of ligands recognised by this receptor (Werling and Jungi 2003).

The main ligand for Toll-like receptor 4 (TLR4) is lipopolysaccharide (LPS), a component of the cell wall of gram negative bacteria. TLR4 also recognises lipoteicoic acid, a component of the cell wall of gram positive bacteria, a number of viral PAMPs and can also recognise endogenous molecules such as HSP60 (heat shock protein 60) (Takeda et al. 2003; Werling and Jungi 2003). White *et al.* (2003) analysed

polymorphism patterns in the bovine TLR4 gene and predicted that positive selection has been acting on the ligand binding domain of the protein. Smirnova *et al.* (2001) reported an excess of rare polymorphisms in in TLR4 gene in human populations and suggested that this is likely due to weak purifying selection (i.e. the removal of deleterious mutations).

The ligand that binds to Toll-like receptor 5 (TLR5) is flagellin. Flagellin is the structural protein from which the flagellae of bacteria are formed.

Toll-like receptor 7 (TLR7) and TLR9 are viral recognition TLRs (vrTLR) and recognise single stranded DNA and CpG DNA, respectively. Single stranded DNA and unmethylated CpG dinucleotides (CpG DNA) are both pathogen associated molecular patterns (PAMPs). A study of vrTLR polymorphism in cattle has recently been reported (Cargill and Womack 2007).

4.1.1.2 Perforin

Perforin is a component of the cytotoxic granules of Natural Killer (NK) cells and cytotoxic T lymphocytes. When a target cell is recognised by the NK cell or the cytotoxic T cell, the granule contents, which include perforin and granzymes, are released into the immunological synapse between the target cell and the NK cell or cytotxic T cell. The end result is the death of the target cell by apoptosis. The mechanism by which this occurs is not fully understood. Early models suggested that perforin multimerizes to form pores in the cell membrane of the target cell through which granzymes enter the target cell and induce apoptosis. The pores on the cell surface can be seen by electron microscopy. A recently proposed model suggested that the role of pores observed in the cell surface act to pump Ca^{2+} into the target cell and trigger membrane repair including the recruitment of lysosome and endosome membranes to the cell membrane. Granzymes, together with perforin, are taken into the target cell by endocytosis and perforin, through an as yet unknown mechanism, perturbs the membrane of the endosome resulting in the release of granzymes into the target cell (Pipkin and Lieberman 2007). Mutations in the perforin gene are associated with FHL (Familial Hemophagocytic lymphohistoscytosis), which is characterised by severe immunodeficiency (Stepp et al. 1999).

4.1.1.3 Mitochondrial Ribosomal Proteins

Ribosomes are subcellular structures whose function is to translate mRNA into protein. There are bacterial ribosomes, mitochondrial ribosomes and cytoplasmic (eukaryotic)

ribosomes. Ribosomes are made up of two subunits, a large subunit and a small subunit, each of which is composed of between one and three rRNA molecules and up to ~49 proteins. Ribosomes and ribosome subunits are named according to their sedimentation coefficient (S) values. These values reflect the sedimentation of the particles in an ultracentrifuge and are not additive. Mammalian mitochondrial ribosomes (55S), like eukaryotic (80S) and bacterial (70S) ribosomes are composed of two subunits, the larger 39S subunit and the smaller 28S subunit. The large subunit consists of a single rRNA molecule 1571 nucleotides in length and 48 proteins, while the 28S subunit consists of a single rRNA molecule of 955 nucleotides and 29 proteins (Mears et al. 2006). This is strikingly different from the ancestral ribosome which is estimated to less than half the number of proteins and more that twice the amount of rRNA (O'Brien 2003). Mammalian mitochondrial ribosomal proteins are encoded by nuclear genes, but function in the mitochondrion while the mitochondrial ribosomal RNAs are encoded by the mitochondrial genome.

Early electrophoretic analyses (Matthews et al. 1978; Pietromonaco et al. 1986) indictated that mitochondrial ribosomal proteins show greater diversity with respect to physical and chemical properties than did their cytoplasmic counterparts leading the authors to speculate that mitochondrial ribosomal proteins are evolving more quickly than cytoplasmic ribosomal proteins. The rate was estimated to be thirteen times greater in MRPs and suggested to be comparable to the rate of evolution of mitochondrial ribosomal RNA. This apparent difference in the rate of evolution is interesting as both mitochondrial and cytoplasmic ribosomal proteins are encoded by nuclear genes.

Analysis of SNP data from bovine ESTs and genome sequencing projects (David Lynn, unpublished) identified a number of MRPs with high levels of polymorphism. To further investigate this, a number of MRPs were sequenced in a panel of *Bos taurus* and *Bos indicus* as well as a number of outgroup species. The three genes sequenced are MRPS5, MRPS14 and MRPL30. MRPS5 and MRPS14 are protein components of the small subunit and both have homologous proteins in the bacterial ribosome. MRPL30 is a component of the large subunit and does not have a homologous protein in the bacterial ribosome.

4.1.1.4 F and TNFAIP8L1: candidate genes on BTA7

Previous work identified a region of reduced microsatellite diversity in African cattle that coincided with a QTL for trypanotolerance (Park 2002). Typanotolerence is a resistance to trypanosomiasis exhibited by a number of African cattle breeds. Two additional

microsatellites were genotyped in the same populations (Ruth Freeman, unpublished). These data were combined with the data from Park (2002). and the microsatellite positions were updated according to the most recent bovine linkage map (Ihara et al. 2004). The region between RM006 (also called CA006) at 25.388cM and BMS1116 at 30.542cM showed lower microsatellite diversity (Figure 4.1). Examination of the genes in this region yielded two interesting candidates, TRIF and TNFAIP8L1, which were selected for resequencing. Both genes are involved in immune response.

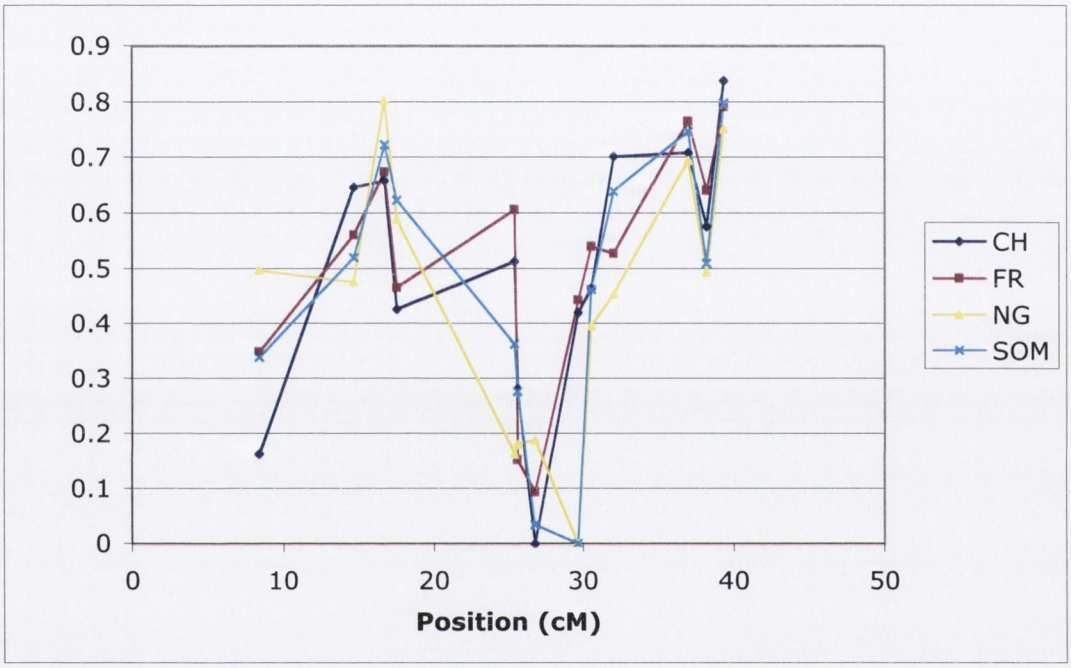


Figure 4.1 Gene Diversity (heterozygosity) across a ~30cM region of BTA7 (updated from Park (2002)).

A region of low microsatellite diversity was identified that was bounded by RM006 at 25.388cM and BMS116 at 30.542cM. The breed abbreviations are as follows: CH, Charolais, FR, Friesian, NG, N'Dama (Guinean) and SOM, Somba.

4.1.1.4.1 TNFAIP8L1

Very little information is available in the literature about TNFAIP8L1 (tumour necrosis factor, alpha induced protein 8-like 1). However, TNFα (tumour necrosis factor alpha), the protein by which TNFAIP8L1 is induced is well studied. It is of particular interest that TNFα , which is an inflammatory cytokine, has been shown to be upregulated during

trypansome infection (Vincendeau and Bouteille 2006). TNFα is a key candidate gene in trypanosomiasis resistance studies in mice, and mice with deficiencies are susceptible to the disease (Iraqi et al. 2001).

4.1.1.4.2 TRIF

TRIF (Toll/IL-1 Receptor Domain Containing Adaptor Protein-Inducting IFNB) or TICAM1 (TIR domain containing adaptor molecule) is one of the five adaptor proteins that facilitate early TLR signalling and as such, play a key role in the innate immune response. TRIF is involved directly in the TLR3 signalling pathway and acts via TRAM in the MyD88 independent pathway of TLR4 signalling (O'Neill and Bowie 2007). The end result of these TRIF mediated signalling pathway is activation of NFκB, activation of interferon regulatory factors (IFRs), or induction of apoptosis via effector proteins that recognise domains on the TRIF protein. Late TLR signalling via TRIF is inhibited by SARM, another of the adaptor proteins.

4.1.1.5 FEZL

Forebrain Embryonic Zinc finger-Like (FEZL), sometimes referred to as ZNF312, is transcription factor that is involved in neuronal development. Sugimoto *et al.* (2006) identified FEZL as a candidate for mastitis resistance by QTL mapping in six half-sib families. Mastitis is an inflammation of the mammary gland occurring as a result of bacterial infection and is a great economic burden on the dairy industry.

They identified a three base pair insertion in the glycine stretch of the protein. This insertion was associated with an increased incidence of mastitis in the half-sib families. FEZL induces expression of SEMA5a (Semaphorin5A), which in turn induces expression of a number of genes involved in immune response, including TNF and IL-8. The cattle with the insertion (designated 13G) showed significantly lower levels of FEZL-induced SEMA5A expression after infection. They also showed that FEZL is induced in the late stage of infection and that this induction occurs via a pathway involving TLR4 and TRIF, both of which have been sequenced as part of this study.

4.1.1.6 The additional data: ART4, CD2, IL2, IL13 and TYROBP

ART4, CD2, IL2, IL13 and TYROBP were identified as candidate genes for positive selection in cattle by inter-genomic comparisons between cattle, pig, human and mouse (Freeman et al. In Prep.; Lynn et al. 2005). Orthologous genes were identified from

assemblies of bovine and porcine ESTs and mRNA sequences and mouse and human sequences. Alignments of four way orthologous gene were subjected to maximum likelihood analysis of the ratio of the rate of nonsynonymous substitutions to the rate synonymous substitutions as implemented in CODEML (PAML). This identified a number of candidate genes which showed a greater than expected ratio of nonsynonymous to synonymous substitutions in the lineage leading to cattle. These candidates were sequenced in (Freeman et al. In Prep.) and the raw data were obtained for analysis here.

4.1.1.6.1 ART4

ART4 is a member of the mono ADP-ribosyltransferase family whose members (ART1-ART5) brings about a posttranslational protein modification involving the transfer of the ADP-ribose moiety from NAD+ to a specific amino acid in the target protein. The amino acid receptor of ART4 has not yet been characterised. ART4 has been shown to be activated in A549 cells (human alveolar basal epithelial cells) after stimulation with LPS, LTA, PAM and PG and it was also shown that TLR2 and TLR4 are essential for activation to occur (Balducci et al. 2007). Human ART4 is an erythrocyte glycoprotein that has been identified as the Dombrock blood group (Gubin et al. 2000). Two major antigens of this blood group (Do^a and Do^b) have significanty different frequencies in human populations and are associated with severe transfusion reactions.

4.1.1.6.2 CD2

CD2 is a cell surface glycoprotein expressed on T-cells and NK cells. It is involved in cell-cell adhesion and facilitates T cell activation through interaction with its counter-receptor, CD58 in non-rodent mammals (Brossay et al. 2003; Selvaraj et al. 1987; Shimojima et al. 2002) and CD48 in rodents (Kato et al. 1992), expressed on the surface of antigen presenting cells (APCs). CD2 and other similar cell surface proteins have low affinity for their counter receptors (Dustin et al. 1997), a fact that may be exploited by viruses, which can evolve a higher affinity for these receptors in order to invade the host cell. CD2 has been shown to be under positive selection in the bovine linage (Anisimova and Yang 2007; Lynn et al. 2005).

4.1.1.6.3 Interleukins (IL2 & IL13)

Interleukins are secreted signalling molecules (cytokines), first identified in leukocyte signalling, which are critical in immune response. IL2 (interleukin 2) is a central

regulator of the immune system involved in processes including the production of cytotoxic T-cells, proliferation of immunoglobulins by B-lymphocyes and induction and maintenance of NK cells (Gaffen and Liu 2004). IL13 (interleukin 13) plays a critical role in allergic inflammation and in susceptibility to parasites such as helminths, schistosomes and nematodes (Bancroft et al. 1998; Gibbs 2005; McKenzie et al. 1998).

4.1.1.6.4 TYROBP

TYROBP (Tyrosine Kinase Binding Protein, also known as DAP12 or KARAP) is an immunoreceptor tyrosine-based activation motif (ITAM) containing adaptor molecule which is expressed on the cell surface as a disulfide-bonded dimer and is involved in the activation of NK cell anti-viral response via 'non-inhibitory' members of the KIR family (Campbell and Colonna 1999; Lanier et al. 1998). On binding to the receptor via a conserved basic amino acid in the transmembrane domain, the ITAM domain is tyrosine phosphorylated leading to the activation of the NK cell. KIRs have been shown to be rapidly evolving in a wide range of species, including cattle, and the KIR gene family has undergone lineage specific expansions or contractions, making the KIR and the molecules with which they interact, interesting subjects for evolutionary analysis.

4.1.2 The Coalescent

The Coalescent is a stochastic process that describes the ancestry of a sample of sequences (sampled at the present time) by tracing each individual sequence backward through time until the most recent common ancestor (MRCA) of all sampled sequences is reached (Hudson 1990; Kingman 1982). The coalescent is based on the Wright Fisher model and assumed the infinite allele model (IAM) of mutation, which is to say that there are no back mutations or recurrent mutations. In addition, a number of other assumptions inherent in the Wright-Fisher model are made. These include no recombination, constant population size, discrete and non-overlapping generations, no structure and equal fitness of all individuals in the population. From the point of view of coalescent analysis, this means that there is equal probability of any two lineages coalescing.

Once the random tree topology had been constructed, the branch lengths are determined by assigning times to the coalescent events. The probability of a coalescent event occurring in a given generation is dependent on the number of lineages in that generation and is given by:

$$P_{coal} = \frac{\binom{n}{2}}{N}$$

where n is the number of lineages and N is the haploid population size. The probability of a coalescent event occurring changes after each such event as the number of lineages has been reduced. This result of this is that the time between coalescent events is exponentially distributed. Finally mutations are added to the genealogy according to a Poisson process. Hudson (1990) showed that changes in N can be incorporated into the coalescent modelling by the introduction of a non-linear scaled factor. Changes in population size do not affect the topology of the genealogy, but affect both the branch length (the timing of the coalescent events) and the positions of the mutations on the genealogy. When the population size is smaller, the probability of two lineages coalescing is greater and when the population is larger, the probability of two lineages coalescing is smaller. A larger population size results in longer branch lengths and since the positioning of the mutations on the tree is dependent on branch lengths, if the ancestral population size was smaller than the present population size, more mutations will be placed on the external branches (before the first coalescent event) resulting in more singletons. Conversely, if the ancestral population size was greater than the present day population size, then the external branch lengths will be shorter and will have fewer mutations resulting in a dearth of singletons in the population.

Coalescent-based analyses are emerging as an extremely useful tool in the analysis of patterns of genetic variation. By using a backwards through time approach, only those lineages that exist at the present time are modelled. This makes the approach computationally effieicent relative to the forward in time approach. As new methods are developed, more sophisticated and more realistic demographic histories can be modelled, for example, see (Antao et al. 2007).

4.2 Materials and Methods

4.2.1 Screening sets

Twelve genes were fully or partially resequenced in one of two sequencing panels.

"Panel A" was made up of 5 European *Bos taurus*, 5 African *Bos taurus*, 5 *Bos indicus* and one outgroup species, *Bos gaurus* (Gaur) (Table 4.1). Admixture analysis (Freeman 2003) of Somba cattle meant that individuals with lowest estimated proportions of zebu ancestry could be selected.

"Panel B" included 20 European *Bos taurus*, representing 10 breeds from across the continent, 11 African *Bos taurus* representing three West African breeds known to have little or no zebu introgression and 11 Asian *Bos indicus* representing 4 Indian breeds. In addition, one *Capra aegagrus hircus* (goat), one *Ovis aries* (sheep), one *Bos gaurus* (Gaur), one *Bison bison* (Plains Bison) and one *Bubalus bubalis* (Water buffalo) were included in the panel. (Table 4.2).

The lineage leading to sheep and goat is estimated to have diverged from the lineage leading to modern cattle in the region of 13.14 – 27.99 million years ago (Gu et al. 2007) based on an analysis of mitochondrial DNA. The lineage leading to water buffalo diverged more recently, at an estimated 10.54 – 13.85 million years ago. This study did not consider Plains Bison or Gaur. Previous work in the lab on Bovidae mitochondrial DNA sequence analysis has indicated that cattle are more closely related to Plains Bison than to Gaur (Finlay 2006), but AFLP analysis has indicated that cattle are more closely related to Gaur than to Plains Bison (Buntjer et al. 2002).

Extracted DNA samples were available in the lab and have previously been described (Freeman et al. In Prep.; MacHugh 1996; Park 2002).

Table 4.1 Breed information for panel A.

Sample ID	Breed	Origin
HAR10b	Hariana	India
SAH7a	Sahiwal	India
SAH8a	Sahiwal	India
THA2a	Tharparker	India
THA2b	Tharparker	India
CH10	Charolais	Europe
JER9	Jersey	Europe
SIM6	Simmental	Europe
FR7	Friesian	Europe
AA4	Aberdeen Angus	Europe
SOM46	Somba (Benin)	Africa
SOM60	Somba (Benin)	Africa
NG6	N'Dama (Guinea)	Africa
NG12	N'Dama (Guinea)	Africa
NG31	N'Dama (Guinea)	Africa
Ga2932	Gaur	Outgroup

Table 4.2 Breed information for panel B.

Sample ID	Breed	Origin
AA4	Aberdeen Angus	Europe
AA5	Aberdeen Angus	Europe
FR6	Friesian	Europe
FR7	Friesian	Europe
NR12	Norwegian Red	Europe
NR31	Norwegian Red	Europe
GB4	German Black	Europe
GB13	German Black	Europe
HL20	Highland	Europe
HL31	Highland	Europe
ALN7	Alentejana	Europe
ALN27	Alentejana	Europe
MRT1	Mertolenga	Europe
MRT17	Mertolenga	Europe
RN8	Romangola	Europe
RN10	Romangola	Europe
SK2	Sikias	Europe
SK4	Sikias	Europe
AB2	Anatolian Black	Europe
AB14	Anatolian Black	Europe
NG6	N'Dama (Guinea)	Africa
NG37	N'Dama (Guinea)	Africa
NG45	N'Dama (Guinea)	Africa
NG49	N'Dama (Guinea)	Africa
S3213	Somba (Togo)	Africa
S3215	Somba (Togo)	Africa
S3217	Somba (Togo)	Africa
S3221	Somba (Togo)	Africa
L17	Lagune	Africa
L19	Lagune	Africa
L38	Lagune	Africa
HAR5a	Hariana	India
HAR8a	Hariana	India
HAR10a	Hariana	India
THR3b	Tharparker	India
THR8b	Tharparker	India
THR10b	Tharparker	India
SAH6a	Sahiwal	India
SAH10a	Sahiwal	India
SAH11a	Sahiwal	India
ONG6	Ongole	India
ONG26	Ongole	India
KH36	Water buffalo	Outgroup
PB3311	Plains Bison 3311	Outgroup
Ga2932	Gaur	Outgroup
Goat	Goat	Outgroup
Sheep Togo	Sheep (Togo)	Outgroup

4.2.2 Primer design for resequencing

Primers were designed using Primer Select or Primer3. In the case of TLR2, TLR4, TLR5, TLR7, TLR9 and PRF1, primers were designed to amplify as much coding sequence as possible in a single fragment. In the case of the remaining genes, the sequencing strategy aimed to cover the whole coding sequence, approximately 1kb of intronic sequence as well as 1kb of flanking sequence at the 5' and 3' end of the gene.

The 'Overlapping Primers' option of The PCR Suite (van Baren and Heutink 2004) was used to design multiple overlapping pairs of primers to amplify FEZL, TRIF and TNFAIP8L1. The PCR Suite is an interface to Primer3. Primers were ordered from MWG (http://www.mwg-biotech.com/) or VH Bio (http://www.vhbio.com/), formerly TAGN. Primer sequences are outlined in Table 4.3.

4.2.3 PCR and Sequencing

The following genes were amplified in the 'Panel A' screening set:

TLR2, TLR4, TLR5, TLR7, TLR9 and PRF1 and sequencing was carried out using Macrogen. The remaining genes were sequenced in the 'Panel B' and sequencing was carried out by AGOWA. Initial optimization of PCR conditions was carried out by running a temperature gradient on the MJ research DYAD PCR machine where T_a ranged from 50 to 65°, such that a different annealing temperature is used for replicate samples. Initially, each temperature gradient was carried out with a final $MgCl_2$ concentration of 1.5mM and 2.0mM. $MgCl_2$ concentration was adjusted on subsequent rounds of optimisation where necessary.

All PCRs were carried out in 96 well plates. The total reaction volume was either 50µl, for those samples sent to Macrogen, or 15µl for those sent to AGOWA. For the samples sent to AGOWA, two amplifications were carried out for each DNA sample and primer pair, such that the forward and reverse sequencing reactions were carried out on different replicates of the PCR product. For those sent to Macrogen, the forward and reverse sequencing reactions were carried out on the same PCR product. Reagent concentrations in the final reaction volume for the PCR were: 1x PCR buffer, 0.2mM of each dNTP, 0.035U/µl of Taq polymerase, ~30ng/µl DNA. Primer concentrations and $MgCl_2$ concentrations varied and are detalied in Table 4.3. The cycling conditions for amplification were: 95° for 15 minutes followed by 35 ° cycles of 95 ° for 30 seconds, T_a for 45 seconds and 72 ° for 1 minute, followed by a final extension step of 72 ° for 15 minutes. Where a touchdown PCR was used, the cycling conditions were as described above, but T_a was reduced by 0.1° each cycle.

Table 4.3 Primer sequences and concentrations, $MgCl_2$ concentrations, annealing temperatures.

Gene	Primer Name	Primer Sequence 5'-3'	length	$MgCl_2$ conc. (mM)	T_a (°C)	Primer conc (μM)	Notes:
FEZL	FEZL_1F	GCCCATGTCATTCCACTTC	19	2.0	61.6	0.3	
	FEZL_1R	AACTGAGCAGCGTCTTGGAC	20				
	FEZL_2F	AGGGCAAGAAATTGCTCAAC	20	2.0	63.8	0.3	
	FEZL_2R	AAAGGAAACTGAGGCCCAAC	20				
	FEZL_3F	CTGGACGGCCCCTTTCTC	18	1.5	58.9	0.3	
	FEZL_3R	ATTCACAGACGAAGGGCTTG	20				
	FEZL_4F	CGCATAAATGCAACCAGTGC	20	2.0	63.8	0.3	
	FEZL_4R	GGGCTCTCACTACAGGATGC	20				
	FEZL_5F	TGACGCGTACGAAGTTTGAC	20	2.0	58.9	0.3	
	FEZL_5R	CTCTTCCGGGACGTTCAC	18				
	FEZL_6F	GCCGAAACGTATTTAATTCCAG	22	2.0	62.7	0.3	
	FEZL_6R	CCAGGAATGACTGAGTGGAAG	21				
	FEZL_7F	CGGATCTCCAAGTGGTTAGG	20	1.5	62.7	0.3	
	FEZL_7R	CGGAGGAGAAGCCAGAAGC	19				
MRPL30	MRPL30_1F	TGTGTGCTCAGTCTTGTCCA	20	2.0	65.0	0.3	touchdown
	MRPL30_1R	TTCCAGACTTCAGAAGGTGTGAT	23				
	MRPL30_2F	TTTTGGGGGTACATTGAAGC	20	2.0	60.0	0.3	touchdown
	MRPL30_2R	GGGGTGGGGAGTGAATTAAA	20				

Gene	Primer Name	Primer Sequence 5'-3'	length	$MgCl_2$ conc. (mM)	Ta (°C)	Primer conc (µM)	Notes:
MRPS05	MRPS05_1F	TTCCTTAGGTGTTAAAAATGTTCC	24	2.5	55.0	0.3	touchdown
	MRPS05_1R	TCAGATCTTAGAAGAAAAGGTTGTT	25				
	MRPS05_2F	TAATTTGTGGCGTTGGCAGT	20	2.0	57.0	0.3	touchdown
	MRPS05_2R	GAAGAAAAATTAAGATACCAAAGCA	25				
	MRPS05_3F	CTGCAATGATGCAGATGACC	20	2.0	62.5	0.3	touchdown
	MRPS05_3R	AGCCATTTGGCTGGACTCTA	20				
	MRPS05_4.5F	AAGCTCCCACCAGCAGTATC	20	2.0	64.5	0.3	touchdown
	MRPS05_4.5R	CAGTGGGGGAGATGACTTTG	20				
	MRPS05_6.7F	TCAGTGATGGTGCCAGTGAT	20	2.0	60.0	0.3	touchdown
	MRPS05_6.7R	CCCACGGACAAAATTCTTCA	20				
MRPS14	MRPS14_1F	GTCACTGTCCCCTGGTGTTT	20	2.0	57.0	1.0	0.1µl taq
	MRPS14_1R	TGTTGACCAAGAAAACGAGATT	22				touchdown
	MRPS14_2F	AGCTGATCATGGGCAACTG	19	2.0	57.0	0.3	touchdown
	MRPS14_2R	TCGAACACTTACATGTTGCTAAAA	24				
PRF1	PRF1-45F	CCTGGCTGTCGAGGCTGAGGTC	22	1.5	68.0	0.3	
	PRF1-1276R	GGGCTGCGAGGGAATGGTGTAA	22				
TLR2	TLR2-548F	TCCTGGTTCAGGTCCCTTTATGTCTT	26	1.75	67.7	0.3	
	TLR2-1787R	CGGCCTCCAACGTCTTCAGTT	21				
TLR4	TLR4-596F	ATCAAATGCCCCTACTCAACCT	22	1.5	67.7	0.3	
	TLR4-1951R	GGACCCCTACCACAGACACCAG	22				
TLR5	TLR5-36F	TCTGCTGTTCAAGGACCCCATCAAG	25	2.0	66.9	0.3	
	TLR5-463R	GCCACCTCAAGTACTGCCGTTTCTG	25				
TLR7	TLF7-34F	CAGGTGCTTTCCAGTTGCGACATCT	25	1.75	68.0	0.3	
	TLR7-906R	TGGCCAGGGCATTTTTCAGACAC	23				

Gene	Primer Name	Primer Sequence 5'-3'	length	$MgCl_2$ conc. (mM)	Ta (°C)	Primer conc (µM)	Notes:
TLR9	TLR9-1951F	CAGATGTGGGCGGAGGGAGACC	22	12.5	66.9	0.3	
	TLR9-3029R	GGGCGCAGGATCACCAGCACTAC	23				
TNFAIP8L1	TNFAIP8L1_1F	CCCTGGTGGAGCAGTTTAAG	20	1.5	58.9	0.3	F seq primer was TNFAIP8L1_7F
	TNFAIP8L1_1R	CACGTACCCACACCACAGAG	20				
	TNFAIP8L1_2F	ACTTGCTGTGTTTGGTGCAG	20	1.5	63.0	0.3	touchdown
	TNFAIP8L1_2R	GCCAGAGTCAGACAGGAAGG	20				
	TNFAIP8L1_3F	GATGCTGGATGAGGACAGC	19	2.0	63.0	0.3	touchdown
	TNFAIP8L1_3R	CTCTCTCCCAGCTTTTGGTG	20				
	TNFAIP8L1_4F	TGAAAATTGGGTGATTGCTG	20	2.0	62.7	0.3	
	TNFAIP8L1_4R	TTCCAACCAACAGATTGCAG	20				
	TNFAIP8L1_6F	ACCACCCTGCTCGAGTGC	18	2.0	65.0	0.3	
	TNFAIP8L1_6R	CAAGTCATGGTGGATTAGGG	20				
	TNFAIP8L1_7F	CCAATGCAGGAGACACAGG	19				F sequencing primer for TNFAIP8L1_1

Gene	Primer Name	Primer Sequence 5'-3'	length	$MgCl_2$ conc. (mM)	Ta (°C)	Primer conc (µM)	Notes:
TRIF	TRIF_1F	TAGCCCTAGCCTCTGGAACC	20	1.75	65.0	0.3	poor sequence; replaced by TRIF_0
	TRIF_1R	GGGGCTGTAGAACTGACCTG	20				
	TRIF_2F	GGGGTGACGTCTGTACTGC	19	1.75	65.0	0.3	
	TRIF_2R	AGTGGATCTCAGGGAACACG	20				
	TRIF_3F	GTCCTCAGCCTCACTTTCC	19	2.0	58.9	0.3	
	TRIF_3R	AATTTCTGCTCGGACGACTC	20				
	TRIF_5F	AGGTTGCTTTGGCGAACTAC	20	2.0	63.8	0.3	poor sequence; replaced by TRIF_9F/5R
	TRIF_5R	TTCTGGACACCTCCAGTTCC	20				
	TRIF_7F	GACCCCTTCTAAGGCTCACC	20	2.0	61.6	0.3	
	TRIF_7R	AGGGTGGGAAAGGATGATG	19				
	TRIF_8F	GGGATTCTCCAGGCAAGAAC	20	2.0	61.6	0.3	
	TRIF_8R	TGGCCTCATAAACCCATTTC	20				
	TRIF_9F	ACTCGGCTTACGTCCATAGC	20	1.5	62.7	0.5	TRIF_9F replaced TRIF_5F
	TRIF_5R	as above					
	TRIF_0F	TAGCCCTAGCCTCTGGAACC	20	2.0	63.0	0.3	TRIF_0 replaced TRIF_1
	TRIF_0R	CACCAAAACCCAGCAAACC	19				

For those fragments sequenced by Macrogen, (TLR2, TLR4, TLR5, TLR7, TLR9 and PRF1) PCR was carried out in a total reaction volume of 50µl.

TLR2, 4, 5, 7, 9 and PRF1 fragments were cleaned up using the JETQUICK PCR purification coulmns (Genomed, Loehne, Germany) according to the manufacturers instructions:

1. PCR product was tranferred to a labelled 1.5ml tube. 400µl of solution H1 was added to tube and was mixed throughly.

2. The prepared mixture of H1 buffer and PCR product was transferred to a JETQUICK spin column that had been prepared by insertion into a 2ml receiver tube. The column was centrifuged in a benchtop microfuge at maximum speed for 1 minute. The flow through was discarded.

3. The spin column was reinserted into the receiver tube and 500µl of solution H2 was added to the spin column. The column was centrifuged in a benchtop microfuge at maximum speed for 1 minute and the flow through was discarded. The column was centrifuged again at maximum speed for 1 minute to removed residual solution.

4. The receiving tube was discarded and the column was placed in a labelled 2ml tube. 50µl of 1xTE was added to the centre of the silica matrix inside the spin column. The column was centrifuged at maximum speed for 2 minutes to ensure maximum possible elution of DNA.

Purified PCR products were transferred to 2ml screw-cap tubes and sent to Macrogen (Seoul, Korea) for sequencing. Sequencing was carried out in the forward and reverse directions using the same primers as were used for PCR. Data were returned in the form of .abi files.

After confirmation of successful amplification by visualisation on agarose gel, the PCR products from MRPS05, MRPS14, MRPL30, FEZL, TRIF and TNFAIP8L1 PCR products were prepared for shipping by desiccation in a speed vac. PCR clean up and sequencing were carried out by AGOWA (Berlin, Germany). Sequencing primers were the same as the PCR primers except where noted (Table 4.3). Data were returned as .abi or .ab1 files.

4.2.4 Additional Data

In addition to the twelve genes described above, data were available for five genes fully resequenced in 'Panel B' (Freeman et al. In Prep.). These genes are: CD2, IL2, IL13,

ART4, and TYROBP. The raw data were obtained and analysed according to the same protocol as the other twelve genes.

4.2.5 Sequence Analysis Pipeline

The sequence analysis pipeline is outlined in Figure 4.1. Sequence data were assembled and analysed using the Phred, Phrap and Consed software (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). Sequences were assembled at the maximum force level (10) due to the inclusion of outgroup sequences. Polymorphic sites and fixed differences between the cattle sequences and the outgroup sequences were identified using PolyPhred (Nickerson et al. 1997; Stephens et al. 2006). All potential polymorphic sites were verified individually and high quality discrepancies were inspected.

The polyphred output was parsed using a python script whose function was to extract genotype data for each individual at each polymorphic site and where necessary, to replace automatic base calls generated by polyphred with manual basecalls such that any edits were incorporated into a single file. This script also extracted the coverage information so that regions of low coverage could be excluded from further analysis.

The genotype data were double checked for discrepancies such as conflicting genotypes at the same site and was formatted for PHASE using a perl script (provided by Brian McEvoy).

PHASE v2.1.1 (Stephens and Scheet 2005) was used to infer haplotypes from the sequence data. With the increasing availability of genetic polymorphism data, most of which is unphased, improvements in statistical methods of haplotype inference are of great interest. Haplotype data contains more information than genotype data as a haplotype describes a number of variants that are inherited together. PHASE v2.1.1 was found to outperform alternative methods of haplotype inference for autosomal data, although the improvement over other methods was small for X chromosome data (Stephens and Scheet 2005). The same study found PHASE v2.1.1 to be superior to other available methods at estimating missing genotypes. Furthermore, it was not sensitive to population structure; analysis of a combined population of African-Americans and European-Americans yielded better results than separate analysis of each population (Stephens and Scheet 2005)..

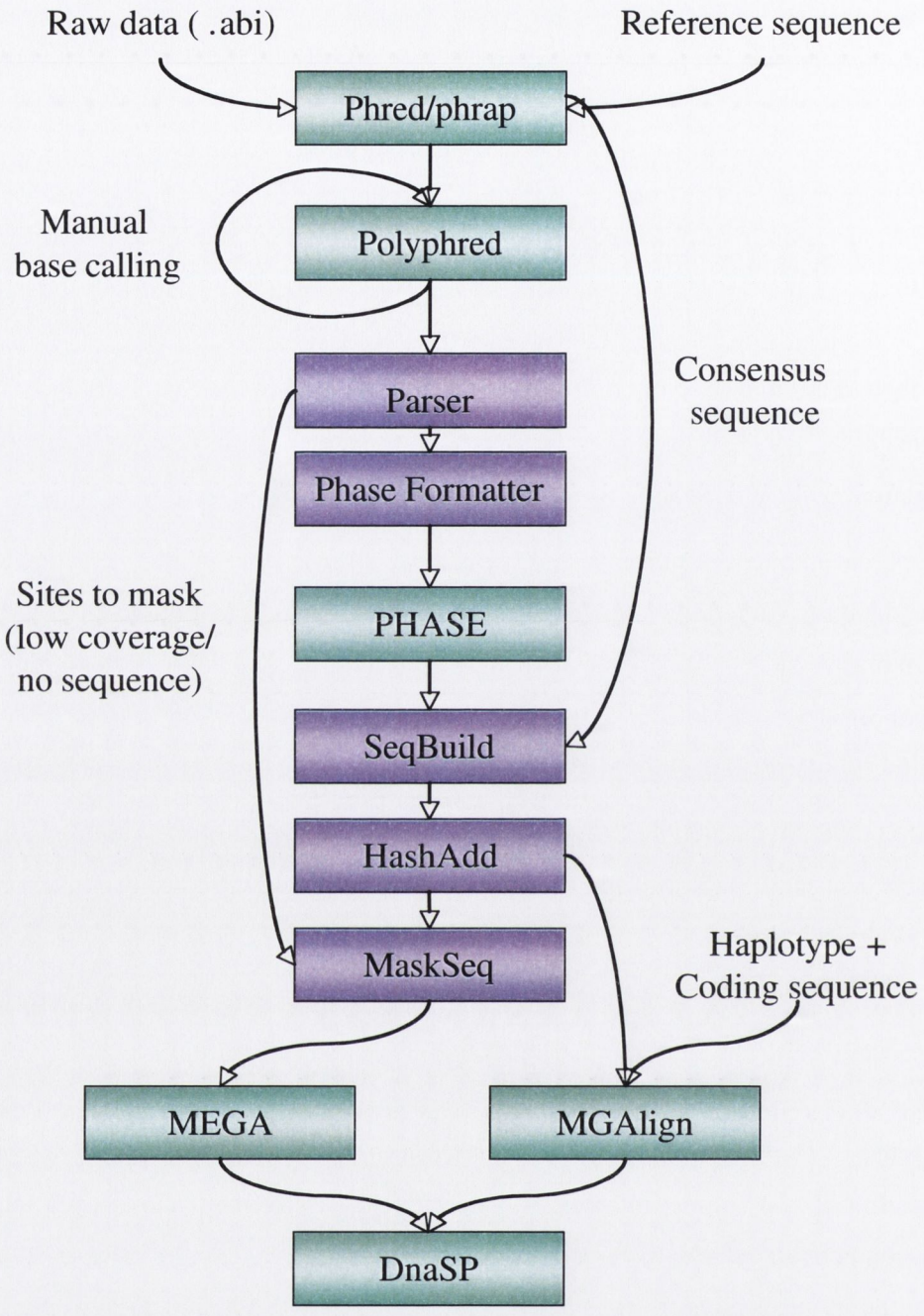


Figure 4.1 Overview of the sequence analysis pipeline used in this study.
Blue boxes indicate publically available software. Purple boxes indicate perl or python scripts written in the lab as follows: Parser (Caitriona Murray), Phase Formatter (Brian McEvoy), SeqBuild, HashAdd and MaskSeq (David Lynn and Tim Downing).

The approach taken here was to analyse all cattle sequences together. A number of DNA samples showed consistently poor sequences, possibly as a result of DNA quality. To

exclude these, a threshold was set at 25% such that sequences with genotype information missing from greater than 25% of sites were excluded.

The Phase output was parsed in Excel making use of the Sequence Toolkit (Park 2002) Macros. Full length haplotypes were reconstructed using the consensus sequence from each Phred/phrap assembly and low coverage regions were masked out using a series of perl scripts (provided by Tim Downing and David Lynn). MGAlign (Lee et al. 2003) was used to align the complete coding sequence to the unmasked haplotype for each gene in order to determine the positions of protein coding sequence. The end result of the sequence analysis pipeline was a file of haplotypes in FASTA format. This file was converted into MEGA (Kumar et al. 2004) format using MEGA and was then ready for analysis in DnaSP (Rozas et al. 2003).

4.2.6 Networks

Reduced Median (RM) networks were created with Network 4.2.0.1 (www.fluxus-engineering.com) using Roehl files generated by DnaSP. No networks were created for TLR5, in which only one polymorphic site and two haplotypes were identified in cattle, and TLR7, in which no polymorphic sites were identified in cattle.

4.2.7 DnaSP

For each gene, the following tests were carried out on a) European *Bos taurus*, b) African *Bos taurus* and c) Indian *Bos taurus*. For those tests requiring an outgroup sequence, Plains Bison (Bison bison) was the outgroup sequence for FEZL, TRIF and TNFAIP8L1 and Gaur (*Bos gaurus*) was the outgroup sequence for TLRs and PRF1. Gaur was the only outgroup species sequenced for the TLRs and PRF1 fragments, and although a Gaur sequence was available for the fully resequenced genes, the haplotypes could not be manually resolved due to the presence of multiple heterozygous sites. For this reason the Plains Bison sequence was used as no gene had more than one heterozygous site. No outgroup sequence was available for any MRP sequence and so none of these tests could be performed.

4.2.8 Haplotype Diversity

One of the simplest measures of variation in a population is the number of haplotypes observed. This, however, is highly dependent on the sample size making it useless for comparisons between different sized samples, such as is the case here. Instead, haplotype diversity (*Hd* or *h*) can be considered. Haplotype diversity is analogous to heterozygosity and is described and estimated by the same equations (Nei 1987). Haplotype diversity is

$$h = 1 - \sum_{i=1}^{m} x_i^2$$

where m is the number of haplotypes and x_i is the frequency of the i^{th} haplotype. Haplotype diversity of a population can be estimated from a sample of that population by:

$$\hat{h} = 2n(1 - \sum \hat{x}_i^2)/(2n-1)$$

where n is the sample size, and $\hat{x}_i$ is the sample frequency of the i^{th} haplotype in the sample, given by:

$$\hat{x}_i = \hat{X}_{ii} + \sum_{j \neq i} \hat{X}_{ij}/2$$

where $\hat{X}_{ij}$ is the frequency of haplotype i,haplotype j in the sample.

Haplotype diversity was estimated for each gene and sample population in DnaSP. The significance of haplotype diversity was estimated by coalescent simulation in DnaSP.

4.2.9 Neutrality Tests

A number of standard tests of neutrality were carried out for each fully or partially sequenced gene fragment using DnaSP (Rozas et al. 2003). These neutrality tests are based on comparisons of difference estimators of θ. θ is equal to $4N_e\mu$ where N_e is the effective population size and μ is the mutation rate. Under neutral theory, all of these estimators are expected to be equal. A number of assumptions are implicit in these tests and violations of these assumptions can result in rejection of the null hypothesis even if all mutations are neutral.

4.2.9.1 Tajima's *D*

Tajima's *D* compares the difference between two estimators of θ. π (often referred to as $\hat{k}$ or $\hat{\theta}_\pi$) is the average number of pairwise nucleotide differences between sequences in a sample, while the second estimator is the number of segregating sites corrected for the sample size (also called $\hat{\theta}_w$ or Watterson's θ) (Tajima 1989):

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

where d is the difference between π and $\hat{\theta}_w$:

$$\hat{\theta}_w = \frac{S}{a_1}$$

where:

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

and

$$\pi = \frac{\sum\sum k_{ij}}{\binom{n}{2}}$$

π is mostly influenced by intermediate frequency variants, while $\hat{\theta}_w$ is most influenced by low frequency variants. A negative value of Tajima's D indicates an excess of low frequency variants and a positive value indicates an excess of intermediate frequency variants.

4.2.9.2 Fay and Wu's *H*

Fay and Wu introduced θ_H, an estimate of θ that is weighted by the homozygosity of derived alleles, rather than ancestral alleles (Fay and Wu 2000). Fay and Wu's H is simply the difference between π (or $\hat{\theta}_\pi$) and $\hat{\theta}_H$:

$$H = \hat{\theta}_\pi - \hat{\theta}_H$$

where:

$$\hat{\theta}_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)}$$

and π can be expressed as:

$$\hat{\theta}_\pi = \sum_{i=1}^{n-1} \frac{2S_i i(n-i)}{n(n-1)}$$

where S_i is the number of derived variants found i times in a sample of n chromosomes.

$\hat{\theta}_H$ is most influenced by high frequency alleles, while $\hat{\theta}_\pi$ is most influenced by intermediate frequency alleles. Fay and Wu's H can detect hitchhiking even when the number of polymorphic sites is low, or when the region across which selection is likely to have had an effect is small.

4.2.9.3 Fu and Li's tests

Fu and Li devised a series of tests of neutrality that depend on whether the mutations are external or internal in the genealogy of the sequences (Fu and Li 1993). The number of external mutations (η_e) is expected to be equal to $\theta = 4N_e\mu$ under the neutral theory. D is the normalised difference between η_e and $\eta_i/(a_n\text{-}1)$, both estimators of θ.

$$D = \frac{\eta_i - (a_n - 1)\eta_e}{\sqrt{Var(\eta_i - (a_n - 1)\eta_e)}} = \frac{\eta - a_n\eta_e}{\sqrt{Var(\eta - a_n\eta_e)}}$$

Similarly, F is the normalised difference between η_e and π:

$$F = \frac{\pi - \eta_e}{\sqrt{Var(\pi - \eta_e)}}$$

where η_e is the number of mutations on external branches (as determined from outgroup sequence), η_i is the number of mutations on internal branches, n is the sample size, π is the mean number of pairwise differences (Tajima 1989) and a_n is the number of segregating sites corrected for sample size, n (equal to a_i in 4.2.9.1).

Fu and Li also derived analogous tests, D^* and F^*, that can be used when no outgroup sequence is available:

$$D^* = \frac{\left(\frac{n}{n-1}\right)\eta - a_n\eta_s}{\sqrt{Var\left(\frac{n}{n-1}\right)\eta - a_n\eta_s}}$$

and

$$F^* = \frac{\left(\frac{n}{n-1}\right)\eta_s - \pi}{\sqrt{Var\left(\frac{n}{n-1}\right)\eta_s - \pi}}$$

where n is the sample size, π is the mean number of pairwise differences, η_s is the number of singletons, η is the total number of mutations and a_n is the number of segregating sites corrected for sample size, n.

A negative value of D indicates an excess of mutations in the external branches whereas a positive value of D indicates a deficiency. An excess of mutations in external branches would be expected if the population had undergone an expansion. Migration could introduce rare alleles into the population.

These tests do not strictly test the neutral mutation hypothesis, but rather, they test whether all the mutations in a region of DNA are neutral. For example, even if all rare variants at a locus are deleterious, the neutral mutation hypothesis still holds, but the hypothesis that all the mutations in the region are neutral does not. It has been suggested that Fu and Li tests are best at detecting background selection (Fu 1997), and that if only Fu and Li tests are significant, then background selection is the most likely explanation.

4.2.9.4 Fu's *Fs*

Fu's *Fs* test (Fu 1997), is based on the probability of observing a greater number of alleles in the data set than would be expected under neutrality. Fu's *Fs* is given by:

$$Fs = \ln\left(\frac{S'}{1-S'}\right)$$

where

$$S' = p(k_0 \leq k \mid \theta = \hat{\theta}_\pi) = \sum_{k \geq k_0} \frac{\mid S_k \mid \hat{\theta}_\pi^k}{S_n(\hat{\theta}_\pi)}$$

S' is the probability of having no fewer than k_0 alleles in a random sample provided that $\theta = \pi$.

Fs is a one sided test. A negative value of *Fs* indicates an excess of recent mutations and a large negative value of *Fs* can be considered evidence that the region has not been evolving neutrally. If only Fu's *Fs* is significant, then population expansion and/or hitchhiking (or possibly recombination) are more likely explanations for the apparent deviations from neutrality.

4.2.10 Assessing significance by coalescent simulations

4.2.10.1 Coalescence simulations in DnaSP

DnaSP (Rozas et al. 2003) version 4 provides a tool for coalescence-based analyses of neutrality tests. By generating large numbers of replicate populations under a model of neutral evolution (using the coalescence based approach), it is possible to compare the values observed in the sequence data to those obtained from the simulated data. This approach was taken to generate the probability and 95% confidence intervals for *Hd*, Tajima's *D*, Fu and Li's *D*, *F*, *D** and *F**, Fay and Wu's *H* and Fu's *Fs*. For each analysis, 1000 replicates were generated. A large constant population, infinite-sites model and no recombination were assumed. Analyses were carried out for each gene and each population given the θ value and sample size obtained from the sequence data.

4.2.10.2 Coalescence using MS

MS (Hudson 2002) is program that generates simulated populations under neutral models using a coalescence-based approach. One advantage of MS over the coalescence based approach available as part of DnaSP is that MS allows the user to describe the demographic model under which the samples should be generated. The coalescence-

based approach in DnaSP is limited to a constant population size. A second advantage of MS is the availability of SCANMS (Ardell 2004) (see section 4.2.10.3)

The same protocol was applied for two hypothetical demographic models (Figure 4.2). Under the first model, the SNM (standard neutral model), the population size has been constant at 10^5 throughout the history of the population. Under the second model, the EXP1 (expansion 1) model, the population size was constant at 10^4 until 10000 years ago, at which time it was instantaneously reduced to 10^3 and immediately began to increase exponentially until it reached its current size (10^5) at the present day.

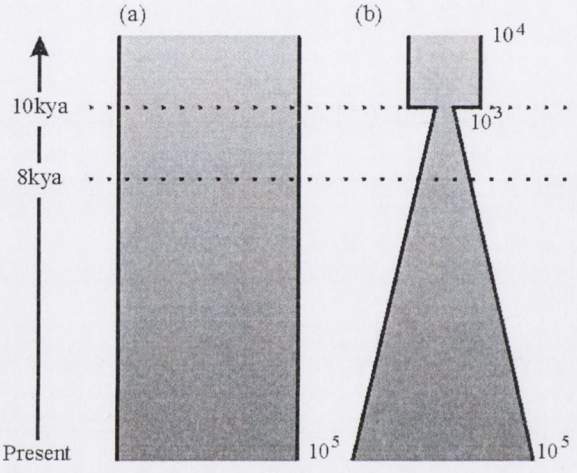


Figure 4.2 Models of demographic history

For each gene in each of the three populations, MS was used to generate 1000 replicate populations each having the same sample size, the same number of segregating sites, and a θ parameter determined by the number of nucleotides that were sequenced in the real data. A mutation rate of 10^{-8} mutations per site per generation and an effective population size of 10^5 were assumed for all populations. "sample-stats", a program that is distributed with MS, was used to calculate the mean values of Tajima's D, θ_H and Fay and Wu's H for each gene in each population from the simulated data. In addition, the standard deviation, the 2.5th percentile, the 5th percentile, the 50th percentile, the 95th percentile and the 97.5th percentile were determined.

4.2.10.3 Sliding window analyses (MS and SCANMS)

SCANMS (Ardell 2004) is a program that simulates a sliding window analysis on the simulated data produced by MS (Hudson 2002). MS was run as described above, but this time the results were piped to SCANMS instead of to sample stats. A window size of

150bp and a slide of 25 bp was used. SCANMS returns the window minimum and maximum values of Tajima's *D*, Fu and Li's *F** and Fu and Li's *D** for each of the replicate populations. From these values, empirical distributions of minimum and maximum values of each statistic were used to determine the significance of the observed values as determined by DnaSP.

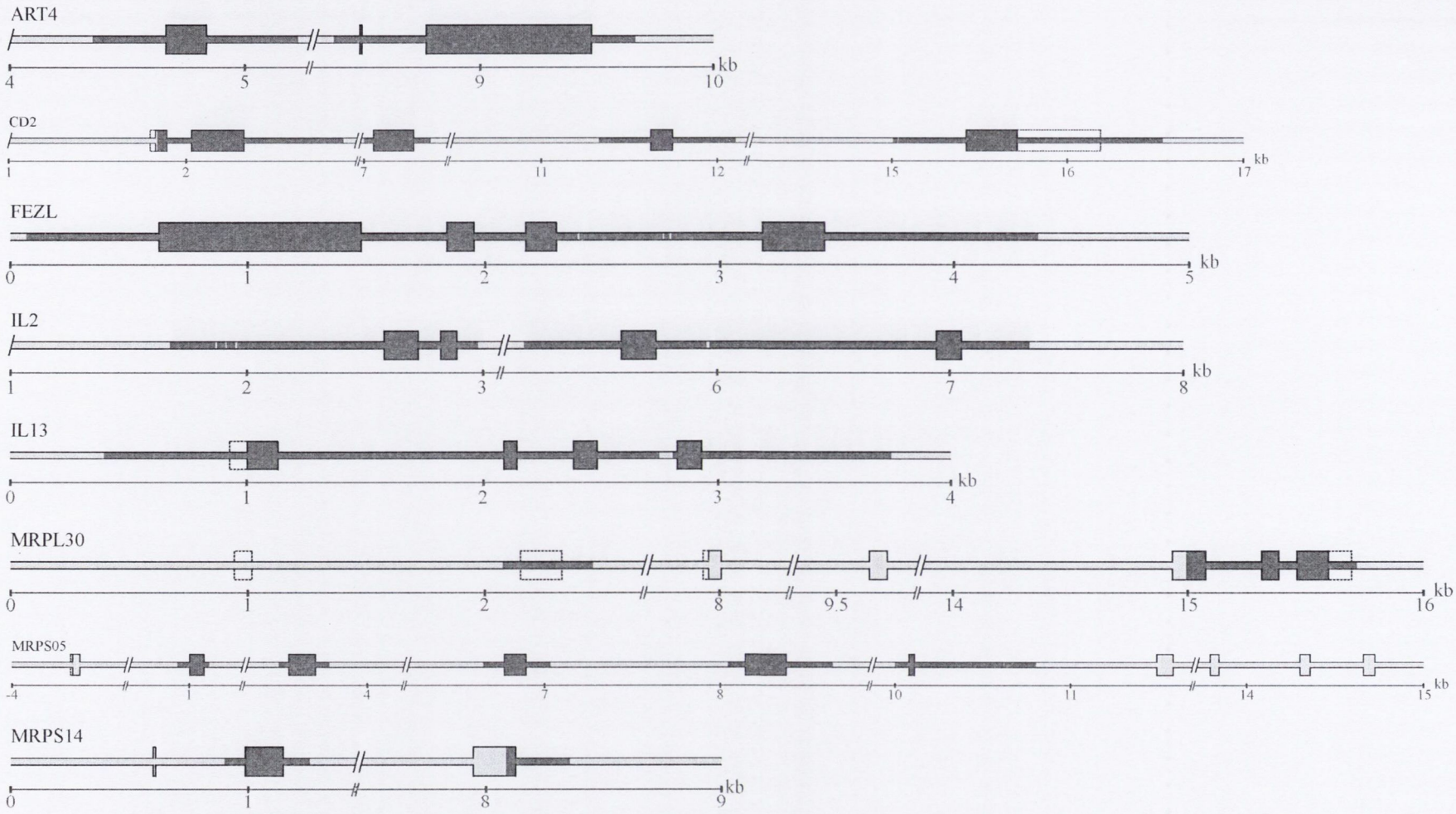
4.3 Results

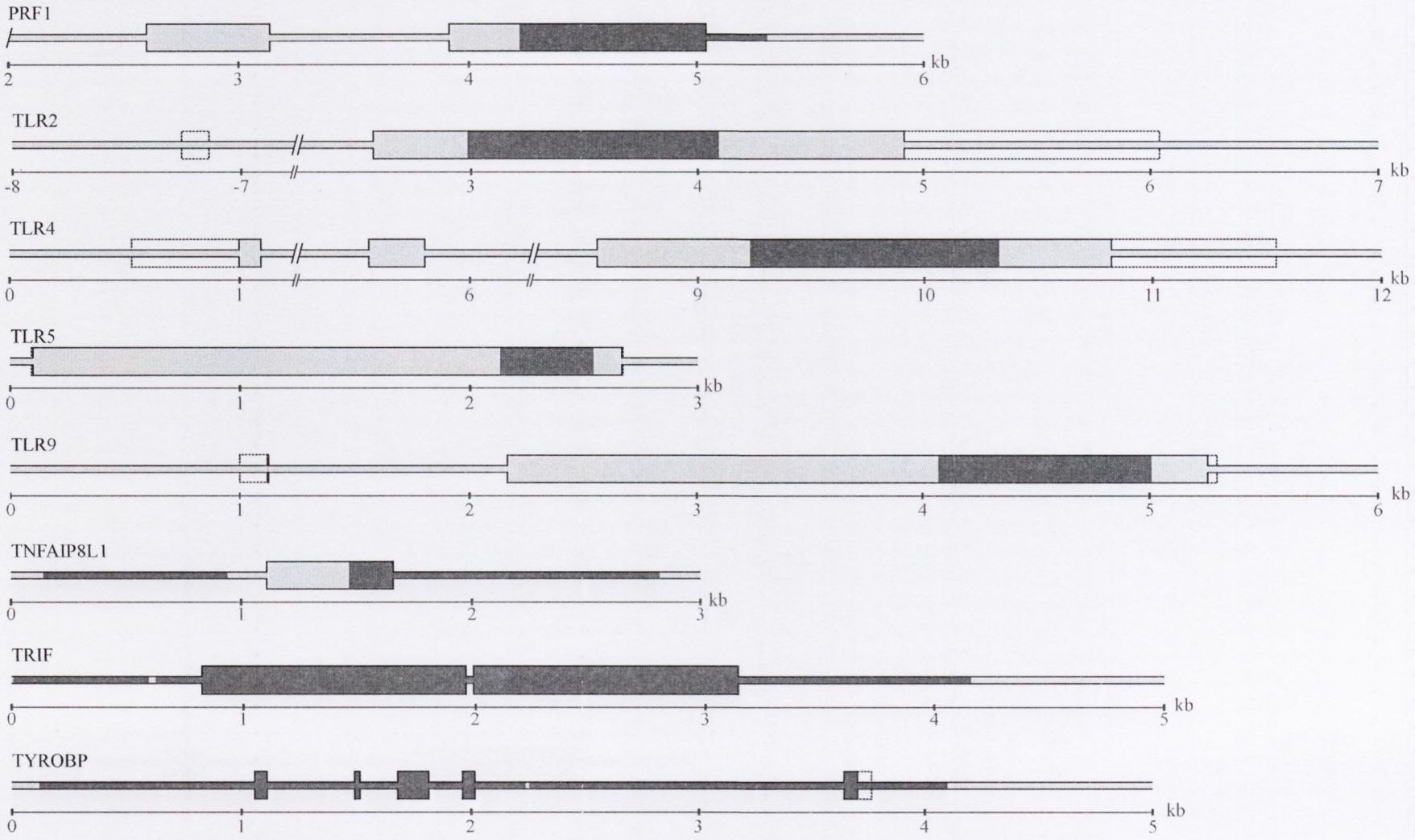
4.3.1 Sequencing

Figure 4.3 shows a schematic diagram for each gene included in the sequencing project. In total 37011bp were sequenced across 17 genes but the number of base pairs of sequence obtained per gene ranged from 412bp (TLR5) to 4282bp (FEZL).

Figure 4.3 Schematic diagram illustrating the structure of each of the genes included in this study and the regions sequenced.

(On following pages.) For each gene the intron/exon structure is indicated with wide rectangles indicating exons and narrow rectangles indicating intronic and intergenic sequences. 5' and 3' UTRs (where described in Ensembl Btau_3.1) are indicated by rectangles with a dotted outline the same height as the exons. Darker shading indicates that this region has been sequenced in the present study. The scale below each figure indicates the Phred/phrap assembly position. Figures drawn using Corel Draw.


ART4
4
5
9
10
kb
CD2
1
2
7
11
12
15
16
17
kb
FEZL
0
1
2
3
4
5
kb
IL2
1
2
3
6
7
8
kb
IL13
0
1
2
3
4
kb
MRPL30
0
1
2
8
9.5
14
15
16
kb
MRPS05
-4
1
4
7
8
10
11
14
15
kb
MRPS14
0
1
8
9
kb


PRF1
kb
2
3
4
5
6
TLR2
kb
-8
-7
3
4
5
6
7
TLR4
kb
0
1
6
9
10
11
12
TLR5
kb
0
1
2
3
TLR9
kb
0
1
2
3
4
5
6
TNFAIP8L1
kb
0
1
2
3
TRIF
kb
0
1
2
3
4
5
TYROBP
kb
0
1
2
3
4
5

4.3.2 Polymorphic sites

The polymorphic sites identified in cattle in this study are outlined in Appendix A.

In total, the 37011bp of sequence yielded 303 SNPs, giving an average frequency of one SNP every 122bp. There was great variation between genes and between populations in terms of the number of SNPs identified with 50 polymorphic sites being identified in TRIF and none in TLR7. As widely differing numbers of sites were sequenced for each region, it is more useful to consider the SNP frequency than the absolute number of SNPs identified. *Bos indicus* showed highest frequency of polymorphic sites while the African *Bos taurus* showed the lowest. In the African *Bos taurus*, on average a SNP was identified every 370bp, in European *Bos taurus* a SNP was identified every 204bp, and in *Bos indicus* a SNP was identified every 167bp.

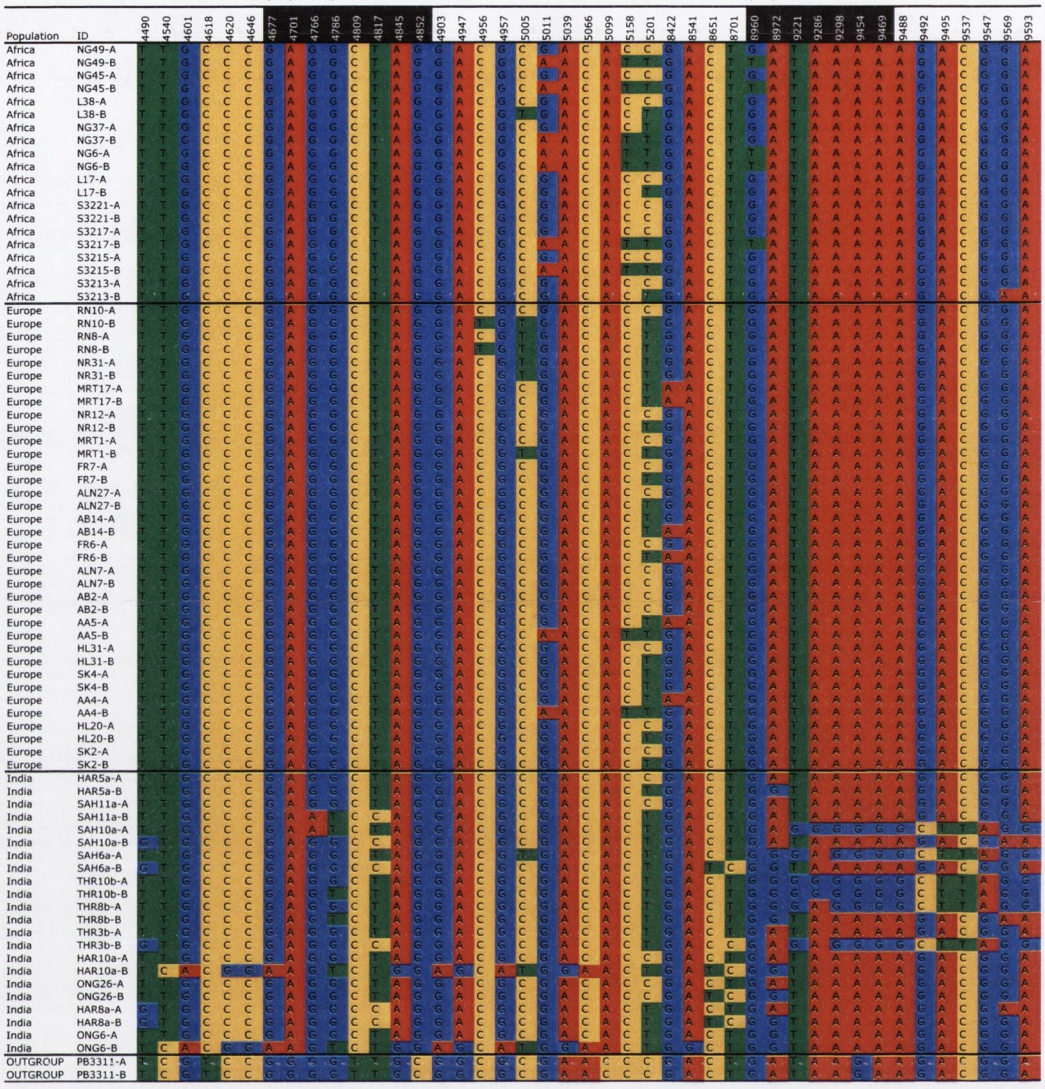
No SNPs were identified in TLR7 (826 bp sequenced) in any population. No SNPs were identified in TLR5 in either of the *Bos taurus* populations (412bp sequenced). Additionally, four other genes showed SNP frequencies of less than one SNP per kb. These were TNFAIP8L1 in African *Bos taurus* (1 SNP every 1081bp), PRF1 in European *Bos taurus* (1 SNP every 1107 bp), FEZL in African *Bos taurus* (1 SNP every 2141 bp). and IL2 in African *Bos taurus* (1 SNP every 3461 bp).

A number of genes showed much higher than average SNP frequencies. The following genes had a SNP frequency of greater than one SNP every 100bp: ART4 in *Bos indicus* (1 SNP every 57bp), TLR4 in *Bos indicus* (1 SNP every 73bp) and MRPS05 in *Bos indicus* (1 SNP every 79bp) and TLR9 in *Bos indicus* (1 SNP every 94bp).

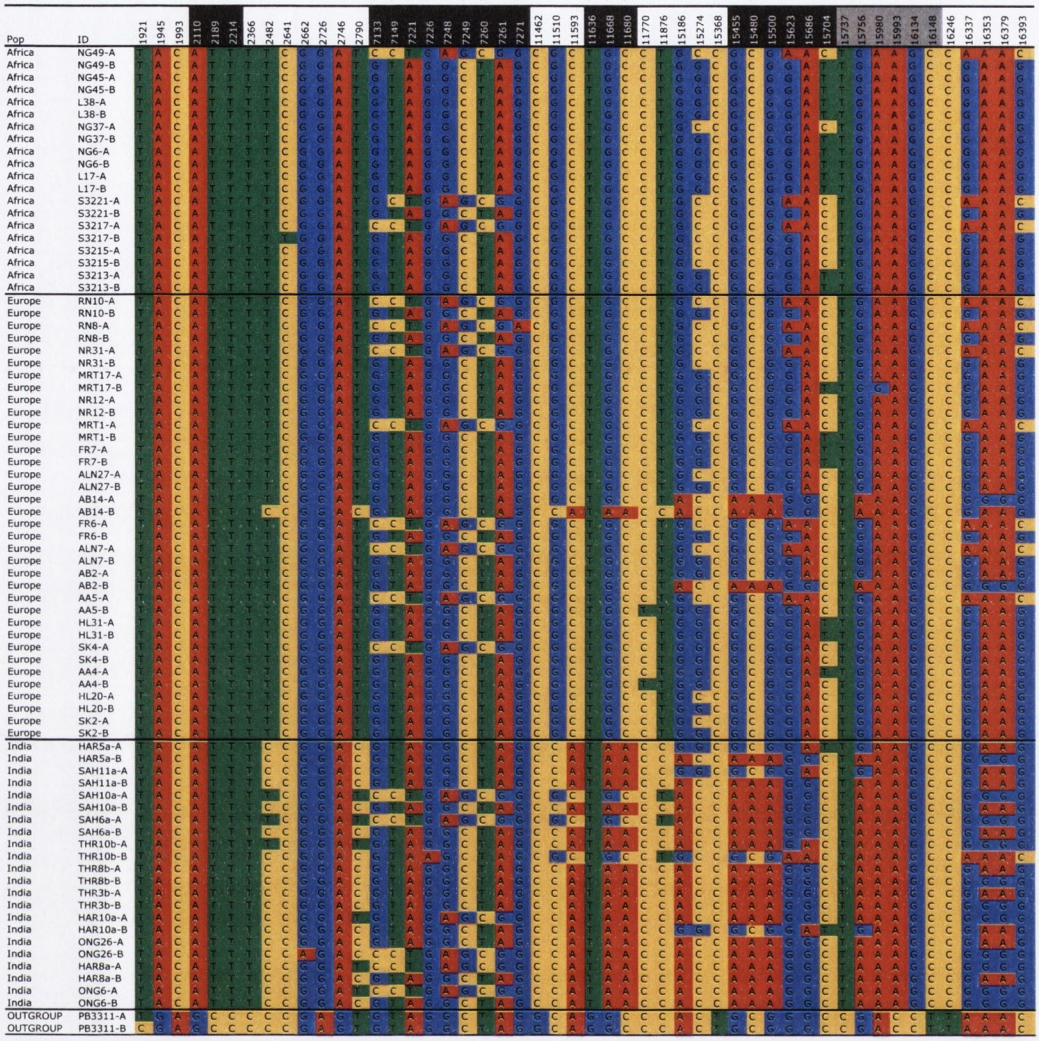
Figure 4.4 Alignments of polymorphic sites.

(Following pages.) The alignments of inferred haplotypes showing polymorphic sites in each cattle population and in the outgroup sequence (Plains Bison or Gaur). Polymorphism position is the assembly position as given in Appendix A. Nonsynonymous substitutions are indicated by *. The assembly positions of polymorphic sites in exons are coloured black, the assembly positions of polymorphic sites in UTRs are coloured grey, and the assembly positions of polymorphic sites in introns or intergenic regions are coloured white. Each animal has two haplotypes, A and B, as inferred using PHASE.

a) ART4


Population
ID
Africa NG49-A
Africa NG49-B
Africa NG45-A
Africa NG45-B
Africa L38-A
Africa L38-B
Africa NG37-A
Africa NG37-B
Africa NG6-A
Africa NG6-B
Africa L17-A
Africa L17-B
Africa S3221-A
Africa S3221-B
Africa S3217-A
Africa S3217-B
Africa S3215-A
Africa S3215-B
Africa S3213-A
Africa S3213-B
Europe RN10-A
Europe RN10-B
Europe RN8-A
Europe RN8-B
Europe NR31-A
Europe NR31-B
Europe MRT17-A
Europe MRT17-B
Europe NR12-A
Europe NR12-B
Europe MRT1-A
Europe MRT1-B
Europe FR7-A
Europe FR7-B
Europe ALN27-A
Europe ALN27-B
Europe AB14-A
Europe AB14-B
Europe FR6-A
Europe FR6-B
Europe ALN7-A
Europe ALN7-B
Europe AB2-A
Europe AB2-B
Europe AA5-A
Europe AA5-B
Europe HL31-A
Europe HL31-B
Europe SK4-A
Europe SK4-B
Europe AA4-A
Europe AA4-B
Europe HL20-A
Europe HL20-B
Europe SK2-A
Europe SK2-B
India HAR5a-A
India HAR5a-B
India SAH11a-A
India SAH11a-B
India SAH10a-A
India SAH10a-B
India SAH6a-A
India SAH6a-B
India THR10b-A
India THR10b-B
India THR8b-A
India THR8b-B
India THR3b-A
India THR3b-B
India HAR10a-A
India HAR10a-B
India ONG26-A
India ONG26-B
India HAR8a-A
India HAR8a-B
India ONG6-A
India ONG6-B
OUTGROUP PB3311-A
OUTGROUP PB3311-B

b) CD2


Pop
ID
1921
1945
1993
2110
2189
2214
2366
2482
2641
2662
2726
2746
2790
7133
7149
7221
7226
7248
7249
7260
7261
7271
11462
11510
11593
11636
11668
11680
11770
11876
15186
15274
15368
15435
15480
15500
15623
15686
15704
15737
15756
15980
15993
16134
16148
16246
16337
16353
16379
16393
Africa NG49-A
Africa NG49-B
Africa NG45-A
Africa NG45-B
Africa L38-A
Africa L38-B
Africa NG37-A
Africa NG37-B
Africa NG6-A
Africa NG6-B
Africa L17-A
Africa L17-B
Africa S3221-A
Africa S3221-B
Africa S3217-A
Africa S3217-B
Africa S3215-A
Africa S3215-B
Africa S3213-A
Africa S3213-B
Europe RN10-A
Europe RN10-B
Europe RN8-A
Europe RN8-B
Europe NR31-A
Europe NR31-B
Europe MRT17-A
Europe MRT17-B
Europe NR12-A
Europe NR12-B
Europe MRT1-A
Europe MRT1-B
Europe FR7-A
Europe FR7-B
Europe ALN27-A
Europe ALN27-B
Europe AB14-A
Europe AB14-B
Europe FR6-A
Europe FR6-B
Europe ALN7-A
Europe ALN7-B
Europe AB2-A
Europe AB2-B
Europe AA5-A
Europe AA5-B
Europe HL31-A
Europe HL31-B
Europe SK4-A
Europe SK4-B
Europe AA4-A
Europe AA4-B
Europe HL20-A
Europe HL20-B
Europe SK2-A
Europe SK2-B
India HAR5a-A
India HAR5a-B
India SAH11a-A
India SAH11a-B
India SAH10a-A
India SAH10a-B
India SAH6a-A
India SAH6a-B
India THR10b-A
India THR10b-B
India THR8b-A
India THR8b-B
India THR3b-A
India THR3b-B
India HAR10a-A
India HAR10a-B
India ONG26-A
India ONG26-B
India HAR8a-A
India HAR8a-B
India ONG6-A
India ONG6-B
OUTGROUP PB3311-A
OUTGROUP PB3311-B

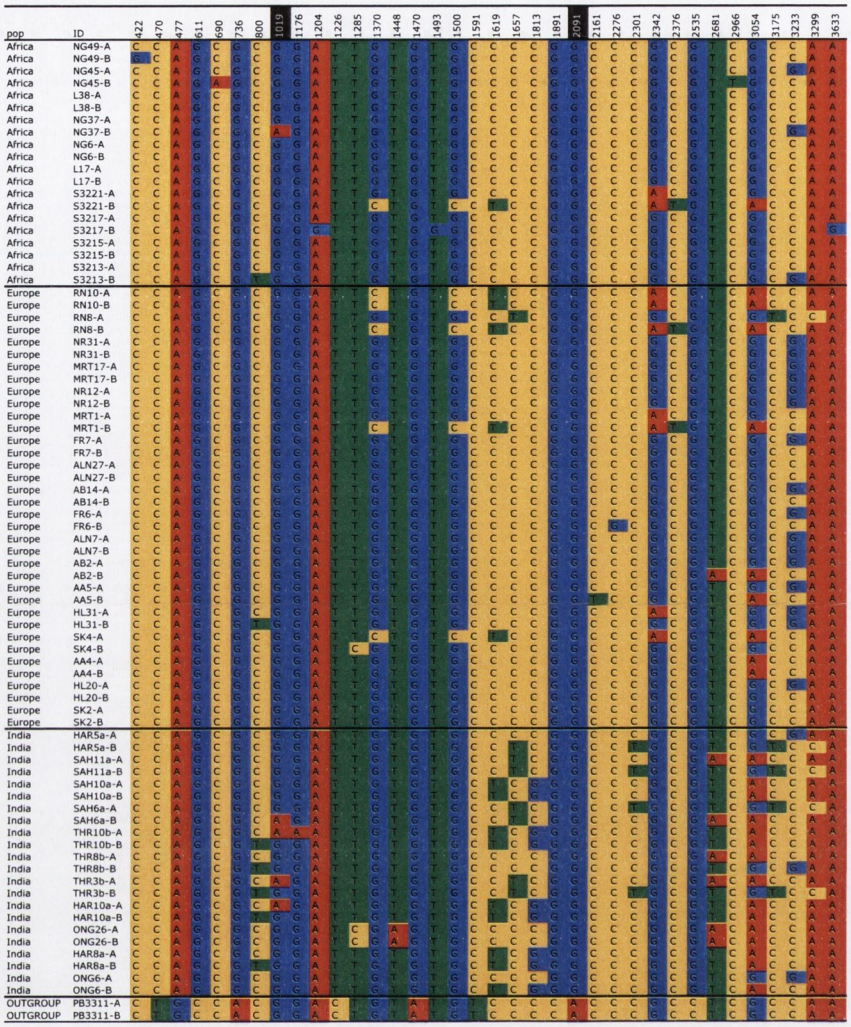
c) FEZL

Pop	ID	476	516	674	873 *	950	961 *	1171 *	1173	1488	1580	1632	1784	1797	1985	1993	2096	2423	2453	2681	2743	2820	2829	2937	3155	3909	4096	4225	4247
Africa	NG49-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG49-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG45-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG45-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	L38-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	L38-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG37-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG37-B	T	G	C	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG6-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	NG6-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3221-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3221-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3217-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3217-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3215-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3215-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3213-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Africa	S3213-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	GB13-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	GB13-B	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	T	G	G	G	C	A	G	G	T	C	C	A	C
Europe	RN10-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	RN10-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	RN8-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	RN8-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	NR31-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	NR31-B	T	G	G	A	A	G	T	G	C	C	G	T	C	G	C	C	G	G	G	C	G	G	G	T	C	C	A	C
Europe	MRT17-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	MRT17-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	NR12-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	NR12-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	MRT1-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	MRT1-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	FR7-A	T	G	G	A	A	G	T	G	C	C	G	T	C	G	C	C	G	G	G	C	G	G	G	T	C	C	A	C
Europe	FR7-B	T	G	G	A	A	G	T	G	C	C	G	T	C	G	C	C	G	G	G	C	G	G	G	T	C	C	A	C
Europe	ALN27-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	ALN27-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	AB14-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	AB14-B	G	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	A	G	C	A	G	G	T	C	C	A	C
Europe	FR6-A	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	FR6-B	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	ALN7-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	ALN7-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	AB2-A	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	AB2-B	T	A	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	AA5-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	AA5-B	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	HL31-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	HL31-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	SK4-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	SK4-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	HL20-A	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	HL20-B	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	SK2-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
Europe	SK2-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	HAR5a-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
India	HAR5a-B	G	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	SAH11a-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
India	SAH11a-B	G	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	SAH10a-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	A	T	C	C	A	C
India	SAH10a-B	T	G	G	A	A	G	C	G	C	C	G	C	C	A	C	C	A	G	C	C	A	G	G	T	T	C	A	C
India	SAH6a-A	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	SAH6a-B	G	G	G	A	A	G	C	A	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	THR10b-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	THR10b-B	G	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	THR8b-A	T	G	G	A	A	G	T	G	C	C	G	C	C	G	C	C	G	G	G	C	A	A	G	T	C	C	A	C
India	THR8b-B	G	G	G	A	A	G	C	A	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	THR3b-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	THR3b-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
India	HAR10a-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	HAR10a-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	ONG26-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
India	ONG26-B	T	G	G	A	A	G	C	G	C	C	G	C	C	A	C	C	A	G	C	C	A	G	G	T	T	C	A	C
India	HAR8a-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	T	A	G	G	T	C	C	A	C
India	HAR8a-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	T	A	G	G	T	C	C	A	C
India	ONG6-A	T	G	G	A	A	G	C	G	C	C	G	C	C	G	C	C	G	G	G	C	A	G	G	T	C	C	A	C
India	ONG6-B	T	G	G	A	A	G	C	G	C	C	G	C	C	G	G	C	G	G	G	C	A	G	G	T	C	C	A	C
OUTGROUP	PB3311-A	T	G	G	G	C	C	C	G	T	G	A	C	A	G	C	C	G	G	G	C	A	G	G	C	C	C	G	G
OUTGROUP	PB3311-B	T	G	G	G	C	C	C	G	T	G	A	C	A	G	C	C	G	G	G	C	A	G	G	C	C	C	G	G

d) IL2

Pop	ID	1760	2381	2541	5283	5344	5391	5710	6031	6058	6133	6135	6150	6167	6300	6342	6400	6708	6750	6775	6887	6941	7057	7156	7170	7242
Africa	NG49-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	NG49-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	NG45-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Africa	NG45-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	L38-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Africa	L38-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	NG37-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Africa	NG37-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	NG6-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	NG6-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	L17-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	L17-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3221-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3221-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3217-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3217-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3215-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3215-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3213-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Africa	S3213-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	RN10-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	RN10-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	RN8-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	T	G	A
Europe	RN8-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	NR31-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	T	A	A
Europe	NR31-B	G	A	T	T	T	C	C	A	G	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	MRT17-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	MRT17-B	T	A	T	T	T	C	C	A	A	A	G	C	C	A	T	A	A	C	C	T	G	C	T	A	A
Europe	NR12-A	G	A	T	T	T	C	C	A	G	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	NR12-B	G	A	T	T	T	C	C	A	G	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	MRT1-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	MRT1-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	FR7-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	FR7-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	ALN27-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	ALN27-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AB14-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AB14-B	G	G	C	T	T	C	C	T	A	A	G	T	T	A	T	A	A	A	T	T	G	C	C	A	A
Europe	FR6-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	FR6-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	ALN7-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	ALN7-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AB2-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AB2-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AA5-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AA5-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	HL31-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	HL31-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	SK4-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	SK4-B	G	A	T	C	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	AA4-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	AA4-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	HL20-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
Europe	HL20-B	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	SK2-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	C	C	T	G	C	C	A	A
Europe	SK2-B	G	A	T	C	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
India	HAR5a-A	G	A	T	T	T	C	C	A	A	A	G	T	C	A	T	A	A	A	C	T	G	C	C	A	A
India	HAR5a-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	SAH11a-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	SAH11a-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	SAH10a-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	SAH10a-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	T	A	T	T	G	C	C	A	A
India	SAH6a-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	SAH6a-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	THR10b-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	THR10b-B	G	G	C	T	T	C	C	T	A	A	G	T	T	A	T	A	A	A	T	T	G	C	C	A	A
India	THR8b-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	THR8b-B	G	G	C	T	T	C	C	T	A	A	G	T	T	A	T	A	A	A	T	T	G	C	C	A	A
India	THR3b-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	THR3b-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	HAR10a-A	G	A	C	T	T	C	C	T	A	A	G	T	T	A	T	A	A	A	T	T	G	C	C	A	A
India	HAR10a-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	ONG26-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	ONG26-B	G	G	C	T	T	C	C	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	HAR8a-A	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	HAR8a-B	G	A	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
India	ONG6-A	G	G	C	T	T	C	C	T	A	A	G	T	T	A	T	A	A	A	T	T	G	C	C	A	A
India	ONG6-B	G	G	C	T	T	C	T	T	A	A	G	T	C	A	T	A	A	A	T	T	G	C	C	A	A
OUTGROUP	PB3311-A	G	A	C	T	G	T	C	T	A	G	A	T	C	G	C	T	A	A	T	T	A	A	C	A	G
OUTGROUP	PB3311-B	G	A	C	T	G	T	C	T	A	G	A	T	C	G	C	T	A	A	T	C	A	A	C	A	G

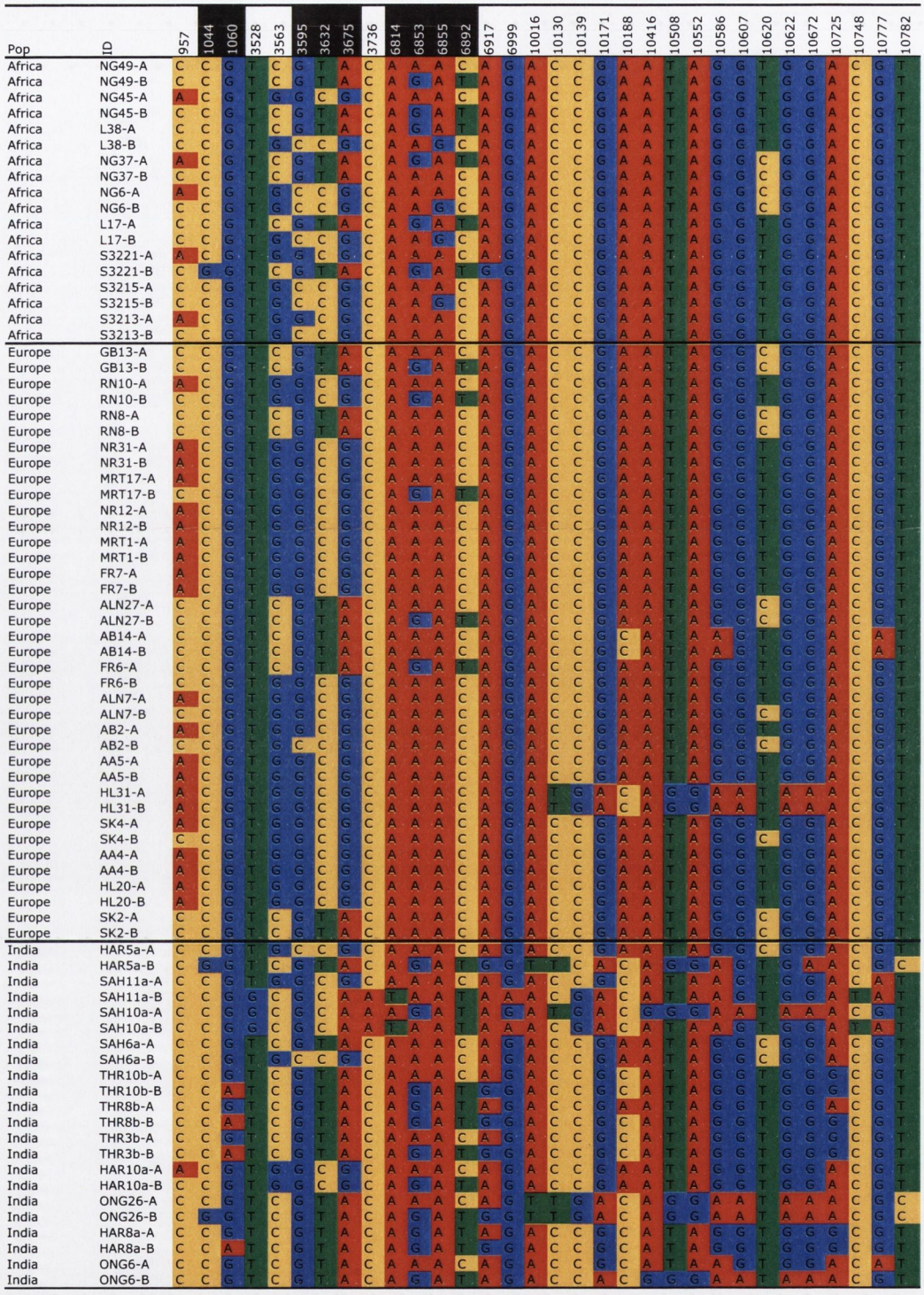
e) IL13


pop
ID
422
470
477
611
690
736
800
1019
1176
1204
1226
1285
1370
1448
1470
1493
1500
1591
1619
1657
1813
1891
2091
2161
2276
2301
2342
2376
2535
2681
2966
3054
3175
3233
3299
3633
Africa NG49-A
Africa NG49-B
Africa NG45-A
Africa NG45-B
Africa L38-A
Africa L38-B
Africa NG37-A
Africa NG37-B
Africa NG6-A
Africa NG6-B
Africa L17-A
Africa L17-B
Africa S3221-A
Africa S3221-B
Africa S3217-A
Africa S3217-B
Africa S3215-A
Africa S3215-B
Africa S3213-A
Africa S3213-B
Europe RN10-A
Europe RN10-B
Europe RN8-A
Europe RN8-B
Europe NR31-A
Europe NR31-B
Europe MRT17-A
Europe MRT17-B
Europe NR12-A
Europe NR12-B
Europe MRT1-A
Europe MRT1-B
Europe FR7-A
Europe FR7-B
Europe ALN27-A
Europe ALN27-B
Europe AB14-A
Europe AB14-B
Europe FR6-A
Europe FR6-B
Europe ALN7-A
Europe ALN7-B
Europe AB2-A
Europe AB2-B
Europe AA5-A
Europe AA5-B
Europe HL31-A
Europe HL31-B
Europe SK4-A
Europe SK4-B
Europe AA4-A
Europe AA4-B
Europe HL20-A
Europe HL20-B
Europe SK2-A
Europe SK2-B
India HAR5a-A
India HAR5a-B
India SAH11a-A
India SAH11a-B
India SAH10a-A
India SAH10a-B
India SAH6a-A
India SAH6a-B
India THR10b-A
India THR10b-B
India THR8b-A
India THR8b-B
India THR3b-A
India THR3b-B
India HAR10a-A
India HAR10a-B
India ONG26-A
India ONG26-B
India HAR8a-A
India HAR8a-B
India ONG6-A
India ONG6-B
OUTGROUP PB3311-A
OUTGROUP PB3311-B

f) MRPL30

Pop	ID	2164	2318	2360	15153	15166	15183	15301	* 15383	15397	15523	15606
Africa	NG49-A	C	G	A	C	T	T	C	A	T	A	A
Africa	NG49-B	C	G	A	C	T	T	C	A	T	A	A
Africa	NG45-A	T	G	A	C	T	T	A	A	T	A	A
Africa	NG45-B	T	G	A	C	T	T	A	A	T	A	A
Africa	NG37-A	C	G	A	C	T	T	C	A	T	A	A
Africa	NG37-B	C	G	A	C	T	T	C	A	T	A	A
Africa	NG6-A	C	G	A	C	T	T	C	A	T	A	A
Africa	NG6-B	C	G	A	C	T	T	C	A	T	A	A
Africa	S3215-A	C	G	A	C	T	T	C	A	T	A	A
Africa	S3215-B	C	G	A	C	T	T	C	A	T	A	A
Africa	S3213-A	C	G	A	C	T	T	C	A	T	A	A
Africa	S3213-B	C	G	A	C	T	T	C	A	T	A	A
Europe	RN10-A	T	G	A	C	T	C	C	T	C	G	G
Europe	RN10-B	T	G	A	C	T	T	C	A	T	G	A
Europe	RN8-A	T	G	A	C	T	C	C	A	C	G	G
Europe	RN8-B	T	G	A	C	T	C	C	A	C	G	G
Europe	NR31-A	C	G	A	C	C	T	C	A	T	A	A
Europe	NR31-B	T	G	A	C	T	T	C	A	T	G	A
Europe	MRT17-A	C	G	A	C	T	T	A	A	T	A	A
Europe	MRT17-B	C	G	A	C	T	T	A	A	T	A	A
Europe	NR12-A	C	G	A	C	T	T	C	A	T	A	A
Europe	NR12-B	T	G	A	C	T	C	C	A	C	G	G
Europe	MRT1-A	C	G	A	C	T	T	C	A	T	A	A
Europe	MRT1-B	C	G	A	C	T	T	C	A	T	A	A
Europe	FR7-A	C	G	A	C	T	T	C	A	T	A	A
Europe	FR7-B	T	G	A	C	T	C	C	T	C	G	G
Europe	ALN27-A	T	G	A	C	T	T	C	A	T	G	A
Europe	ALN27-B	T	G	A	C	T	T	C	A	T	G	A
Europe	AB14-A	T	G	A	C	T	C	C	A	C	G	G
Europe	AB14-B	T	G	A	C	T	C	C	A	C	G	G
Europe	FR6-A	C	G	A	C	T	T	C	A	T	A	A
Europe	FR6-B	T	G	A	C	T	T	C	A	T	G	A
Europe	ALN7-A	C	G	A	C	T	T	C	A	T	A	A
Europe	ALN7-B	C	G	A	C	T	T	C	A	T	A	A
Europe	AB2-A	C	G	A	C	T	T	C	A	T	A	A
Europe	AB2-B	T	G	A	C	T	T	C	A	T	G	A
Europe	AA5-A	C	G	A	C	T	T	C	A	T	A	A
Europe	AA5-B	T	G	A	C	T	T	C	A	T	G	A
Europe	HL31-A	C	G	A	C	T	T	C	A	T	A	A
Europe	HL31-B	C	G	A	C	T	T	C	A	T	A	A
Europe	SK4-A	T	G	A	C	T	C	C	A	C	G	G
Europe	SK4-B	T	G	A	C	T	C	C	A	C	G	G
Europe	AA4-A	C	G	A	T	T	T	C	A	T	A	A
Europe	AA4-B	T	G	A	C	T	T	C	A	T	G	A
Europe	HL20-A	C	G	A	C	T	T	C	A	T	A	A
Europe	HL20-B	C	G	A	C	T	T	C	A	T	A	A
Europe	SK2-A	C	G	A	C	T	T	C	A	T	A	A
Europe	SK2-B	T	G	A	C	T	C	C	A	C	G	G
India	HAR5a-A	T	G	A	C	T	C	C	A	C	G	G
India	HAR5a-B	T	G	A	C	T	T	C	A	T	G	G
India	SAH11a-A	T	G	A	C	T	C	C	A	C	G	G
India	SAH11a-B	T	G	A	C	T	T	C	A	T	G	A
India	SAH10a-A	T	G	A	C	T	C	C	A	C	G	G
India	SAH10a-B	T	G	A	C	T	T	C	A	T	G	A
India	SAH6a-A	T	G	A	C	T	C	C	A	C	G	G
India	SAH6a-B	T	G	A	C	T	C	C	A	C	G	G
India	THR10b-A	C	G	A	C	T	T	C	A	T	A	A
India	THR10b-B	T	G	A	C	T	C	C	A	C	G	G
India	THR8b-A	T	G	A	C	T	C	C	A	C	G	G
India	THR8b-B	T	G	A	C	T	T	C	A	T	G	A
India	THR3b-A	T	C	G	C	T	C	C	A	C	G	G
India	THR3b-B	T	G	A	C	T	C	C	A	C	G	G
India	HAR10a-A	T	G	A	C	T	C	C	A	C	G	G
India	HAR10a-B	T	G	A	C	T	C	C	A	C	G	G
India	ONG26-A	T	G	A	C	T	C	C	A	C	G	G
India	ONG26-B	T	G	A	C	T	C	C	A	C	G	G
India	HAR8a-A	T	G	A	C	T	C	C	A	C	G	G
India	HAR8a-B	T	G	A	C	T	C	C	A	C	G	G
India	ONG6-A	T	G	A	C	T	C	C	A	C	G	G
India	ONG6-B	T	G	A	C	T	C	C	A	C	G	G

g) MRPS05


Pop
ID
957
1044
1060
3528
3563
3595
3632
3675
3736
6814
6853
6855
6892
6917
6999
10016
10130
10139
10171
10188
10416
10508
10552
10586
10607
10620
10622
10672
10725
10748
10777
10782
Africa NG49-A
Africa NG49-B
Africa NG45-A
Africa NG45-B
Africa L38-A
Africa L38-B
Africa NG37-A
Africa NG37-B
Africa NG6-A
Africa NG6-B
Africa L17-A
Africa L17-B
Africa S3221-A
Africa S3221-B
Africa S3215-A
Africa S3215-B
Africa S3213-A
Africa S3213-B
Europe GB13-A
Europe GB13-B
Europe RN10-A
Europe RN10-B
Europe RN8-A
Europe RN8-B
Europe NR31-A
Europe NR31-B
Europe MRT17-A
Europe MRT17-B
Europe NR12-A
Europe NR12-B
Europe MRT1-A
Europe MRT1-B
Europe FR7-A
Europe FR7-B
Europe ALN27-A
Europe ALN27-B
Europe AB14-A
Europe AB14-B
Europe FR6-A
Europe FR6-B
Europe ALN7-A
Europe ALN7-B
Europe AB2-A
Europe AB2-B
Europe AA5-A
Europe AA5-B
Europe HL31-A
Europe HL31-B
Europe SK4-A
Europe SK4-B
Europe AA4-A
Europe AA4-B
Europe HL20-A
Europe HL20-B
Europe SK2-A
Europe SK2-B
India HAR5a-A
India HAR5a-B
India SAH11a-A
India SAH11a-B
India SAH10a-A
India SAH10a-B
India SAH6a-A
India SAH6a-B
India THR10b-A
India THR10b-B
India THR8b-A
India THR8b-B
India THR3b-A
India THR3b-B
India HAR10a-A
India HAR10a-B
India ONG26-A
India ONG26-B
India HAR8a-A
India HAR8a-B
India ONG6-A
India ONG6-B

h) MRPS14

Pop	ID	952	8200
Africa	NG49-A	G	T
Africa	NG49-B	G	C
Africa	NG45-A	A	C
Africa	NG45-B	A	C
Africa	L38-A	G	T
Africa	L38-B	G	T
Africa	NG37-A	A	C
Africa	NG37-B	A	C
Africa	L19-A	G	T
Africa	L19-B	G	T
Africa	NG6-A	G	T
Africa	NG6-B	G	C
Africa	L17-A	G	T
Africa	L17-B	G	T
Africa	S3221-A	G	T
Africa	S3221-B	G	C
Africa	S3217-A	A	T
Africa	S3217-B	A	C
Africa	S3215-A	G	T
Africa	S3215-B	G	C
Africa	S3213-A	G	T
Africa	S3213-B	G	T
Europe	GB13-A	G	T
Europe	GB13-B	G	T
Europe	RN10-A	G	T
Europe	RN10-B	G	T
Europe	GB4-A	G	T
Europe	GB4-B	G	C
Europe	RN8-A	G	T
Europe	RN8-B	G	T
Europe	NR31-A	G	T
Europe	NR31-B	G	T
Europe	MRT17-A	G	T
Europe	MRT17-B	G	T
Europe	NR12-A	G	T
Europe	NR12-B	G	T
Europe	MRT1-A	G	T
Europe	MRT1-B	G	C
Europe	FR7-A	G	T
Europe	FR7-B	G	T
Europe	ALN27-A	G	T
Europe	ALN27-B	G	T
Europe	AB14-A	G	T
Europe	AB14-B	G	C
Europe	FR6-A	G	T
Europe	FR6-B	G	T
Europe	ALN7-A	G	T
Europe	ALN7-B	G	T
Europe	AB2-A	G	T
Europe	AB2-B	G	C
Europe	AA5-A	G	T
Europe	AA5-B	G	T
Europe	HL31-A	G	T
Europe	HL31-B	G	T
Europe	SK4-A	G	T
Europe	SK4-B	G	T
Europe	AA4-A	G	T
Europe	AA4-B	G	T
Europe	HL20-A	G	T
Europe	HL20-B	G	T
Europe	SK2-A	G	T
Europe	SK2-B	G	T
India	HAR5a-A	G	C
India	HAR5a-B	G	C
India	SAH11a-A	G	C
India	SAH11a-B	G	C
India	SAH10a-A	G	C
India	SAH10a-B	G	C
India	SAH6a-A	G	T
India	SAH6a-B	G	C
India	THR10b-A	G	C
India	THR10b-B	G	C
India	THR8b-A	G	C
India	THR8b-B	G	C
India	THR3b-A	G	C
India	THR3b-B	G	C
India	HAR10a-A	G	T
India	HAR10a-B	G	C
India	ONG26-A	G	C
India	ONG26-B	G	C
India	HAR8a-A	G	T
India	HAR8a-B	G	C
India	ONG6-A	G	C
India	ONG6-B	G	C

i) PRF1

Pop	ID	396	570	629	639	683	745	812	1037	1084	1319	5252	5078	5019	5009	4965	4903	4836	4611	4564	4329
Africa	SOM46-A	A	T	C	T	G	T	A	T	C	C										
Africa	SOM46-B	A	T	C	T	G	T	A	T	C	C										
Africa	SOM60-A	C	T	C	T	G	C	G	T	C	C										
Africa	SOM60-B	A	T	C	T	G	T	A	T	C	C										
Africa	NG6-A	A	T	C	T	G	T	A	T	C	C										
Africa	NG6-B	A	T	C	T	G	T	A	T	C	C										
Africa	NG12-A	A	T	C	T	G	T	A	T	C	C										
Africa	NG12-B	A	T	C	T	G	T	A	T	C	C										
Africa	NG31-A	A	T	C	T	G	T	A	T	C	C										
Africa	NG31-B	A	T	C	T	G	T	A	T	C	C										
Europe	CH10-A	A	T	C	T	G	T	A	T	C	T										
Europe	CH10-B	A	T	C	T	G	T	A	T	C	T										
Europe	JER9-A	A	T	C	T	G	T	A	T	C	C										
Europe	JER9-B	A	T	C	T	G	T	A	T	C	T										
Europe	SIM6-A	A	T	C	T	G	T	A	T	C	C										
Europe	SIM6-B	A	T	C	T	G	T	A	T	C	C										
Europe	FR7-A	A	T	C	T	G	T	A	T	C	C										
Europe	FR7-B	A	T	C	T	G	T	A	T	C	C										
Europe	AA4-A	A	T	C	T	G	T	A	T	C	C										
Europe	AA4-B	A	T	C	T	G	T	A	T	C	C										
India	HAR10a-A	C	T	C	T	G	C	A	T	C	C										
India	HAR10a-B	C	T	C	T	G	C	G	T	C	C										
India	SAH7a-A	C	T	C	T	G	C	G	T	C	C										
India	SAH7a-B	C	T	C	T	G	C	G	C	T	C										
India	SAH8a-A	C	T	C	T	G	C	G	C	T	C										
India	SAH8a-B	C	T	C	T	G	C	G	C	T	C										
India	THA2a-A	C	T	C	T	G	C	A	T	C	C										
India	THA2a-B	C	T	C	T	G	C	G	C	T	C										
India	THA3a-A	C	T	C	T	G	C	A	T	C	C										
India	THA3a-B	C	T	C	T	G	C	G	T	C	C										
Outgroup	Ga2932-A	C	C	T	C	A	C	G	T	C	C										
Outgroup	Ga2932-B	C	C	T	C	A	C	G	C	C	C										

j) TLR2

Pop	ID	3019	3166	3195	3245	3426	3450	3509	3518	3541	3542	3574	3593	3778	3814	3872	4018	4068
Africa	SOM46-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	SOM46-B	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	SOM60-A	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	SOM60-B	A	G	G	T	A	A	G	T	A	T	G	A	C	A	A	T	T
Africa	NG6-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	NG6-B	A	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	NG12-A	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	NG12-B	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	NG31-A	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Africa	NG31-B	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	CH10-A	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	CH10-B	A	A	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	JER9-A	G	G	G	T	A	A	T	T	A	A	A	A	C	G	C	C	T
Europe	JER9-B	A	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	SIM6-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	SIM6-B	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	FR7-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	FR7-B	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	AA4-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
Europe	AA4-B	G	G	G	T	A	A	G	T	A	T	G	A	C	A	A	C	T
India	HAR10a-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
India	HAR10a-B	G	G	G	A	A	A	G	T	A	A	A	A	C	G	C	C	T
India	SAH7a-A	G	G	G	A	A	A	G	T	A	A	A	A	C	G	C	C	T
India	SAH7a-B	G	G	G	A	A	A	G	T	A	A	A	A	C	G	C	C	T
India	SAH8a-A	G	G	G	A	A	A	G	T	A	A	A	A	T	G	C	C	G
India	SAH8a-B	G	G	G	A	A	A	G	T	A	A	A	A	T	G	C	C	G
India	THA2a-A	G	G	A	T	A	A	G	T	A	T	G	A	C	A	A	C	T
India	THA2a-B	G	G	G	T	A	A	G	T	A	A	A	C	C	G	C	C	T
India	THA3a-A	G	G	G	T	A	A	G	T	A	A	A	C	C	G	C	C	T
India	THA3a-B	G	G	G	T	A	A	G	T	A	A	A	C	C	G	C	C	T
Outgroup	Ga2932-A	G	G	G	A	T	G	G	C	G	T	A	A	C	G	C	C	T
Outgroup	Ga2932-B	G	G	G	A	T	G	G	C	G	T	A	A	C	G	C	C	T

k) TLR4

Pop	ID	9337	9416	9439	9450	9464	9818	9900	9953	10064	10124	10163	10172	10244	10245	10289	10318	10325
		*		*											*		*	
Africa	SOM46-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	SOM46-B	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	SOM60-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	SOM60-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	C	T
Africa	NG6-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	NG6-B	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	NG12-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	NG12-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	C	T
Africa	NG31-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Africa	NG31-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	T	T
Europe	CH10-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Europe	CH10-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	T	T
Europe	JER9-A	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	T	T
Europe	JER9-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	T	T
Europe	SIM6-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Europe	SIM6-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	T	T
Europe	FR7-A	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	C	T
Europe	FR7-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	T	T
Europe	AA4-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
Europe	AA4-B	C	A	A	T	T	A	C	C	T	T	T	C	G	G	C	C	T
India	HAR10a-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
India	HAR10a-B	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
India	SAH7a-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
India	SAH7a-B	C	G	A	T	G	G	C	T	C	T	T	C	G	G	C	C	C
India	SAH8a-A	C	A	A	T	T	A	C	T	T	T	T	C	G	G	C	C	T
India	SAH8a-B	C	G	A	T	G	G	C	T	C	T	T	C	G	G	C	C	C
India	THA2a-A	C	G	A	T	G	G	C	C	T	T	C	C	A	A	C	C	T
India	THA2a-B	C	G	A	T	G	G	C	C	T	T	C	C	A	A	C	C	T
India	THA3a-A	C	G	G	C	T	G	C	T	T	T	T	C	G	G	C	C	T
India	THA3a-B	A	G	A	T	G	G	C	T	T	C	T	T	A	A	A	C	T
Outgroup	Ga2932-A	C	G	G	C	T	G	T	T	C	T	C	C	A	A	C	C	C
Outgroup	Ga2932-B	C	G	G	C	T	G	T	T	C	T	C	C	A	A	C	C	C

l) TLR5

Pop	ID	2179	2308
Africa	SOM46-A	C	G
Africa	SOM46-B	C	G
Africa	SOM60-A	C	G
Africa	SOM60-B	C	G
Africa	NG6-A	C	G
Africa	NG6-B	C	G
Africa	NG12-A	C	G
Africa	NG12-B	C	G
Africa	NG31-A	C	G
Africa	NG31-B	C	G
Europe	CH10-A	C	G
Europe	CH10-B	C	G
Europe	JER9-A	C	G
Europe	JER9-B	C	G
Europe	SIM6-A	C	G
Europe	SIM6-B	C	G
Europe	FR7-A	C	G
Europe	FR7-B	C	G
Europe	AA4-A	C	G
Europe	AA4-B	C	G
India	HAR10a-A	C	C
India	HAR10a-B	C	C
India	SAH7a-A	C	C
India	SAH7a-B	C	C
India	SAH8a-A	C	C
India	SAH8a-B	C	C
India	THA2a-A	C	G
India	THA2a-B	C	G
India	THA3a-A	C	G
India	THA3a-B	C	G
Outgroup	Ga2932-A	C	C
Outgroup	Ga2932-B	T	C

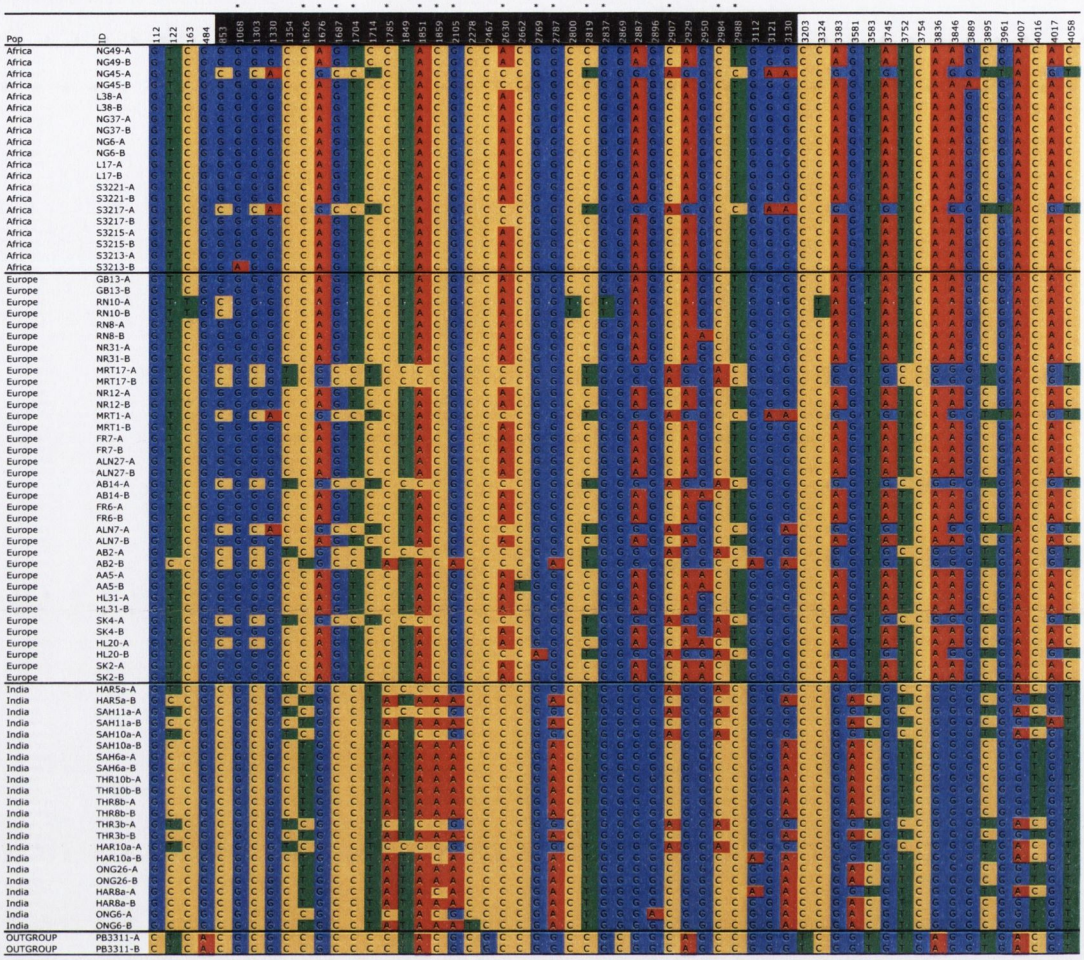
m) TLR9

Pop	ID	4526	4580	4589	4625
Africa	NG6-A	A	T	G	G
Africa	NG6-B	A	C	G	A
Africa	NG12-A	A	T	G	G
Africa	NG12-B	A	C	G	A
Africa	NG31-A	A	T	G	G
Africa	NG31-B	A	C	G	A
Europe	CH10-A	A	T	G	G
Europe	CH10-B	A	T	G	G
Europe	JER9-A	A	T	G	G
Europe	JER9-B	A	T	G	G
Europe	SIM6-A	A	T	G	G
Europe	SIM6-B	A	C	G	G
Europe	AA4-A	A	T	G	G
Europe	AA4-B	A	T	G	G
India	HAR10a-A	A	T	A	G
India	HAR10a-B	A	T	G	G
India	SAH7a-A	A	T	G	G
India	SAH7a-B	A	T	G	G
India	SAH8a-A	A	T	A	G
India	SAH8a-B	A	T	G	G
India	THA2a-A	A	T	A	G
India	THA2a-B	A	T	G	G
India	THA3a-A	A	T	G	G
India	THA3a-B	A	T	G	G
Outgroup	Ga2932-A	A	C	G	G
Outgroup	Ga2932-B	G	C	G	G

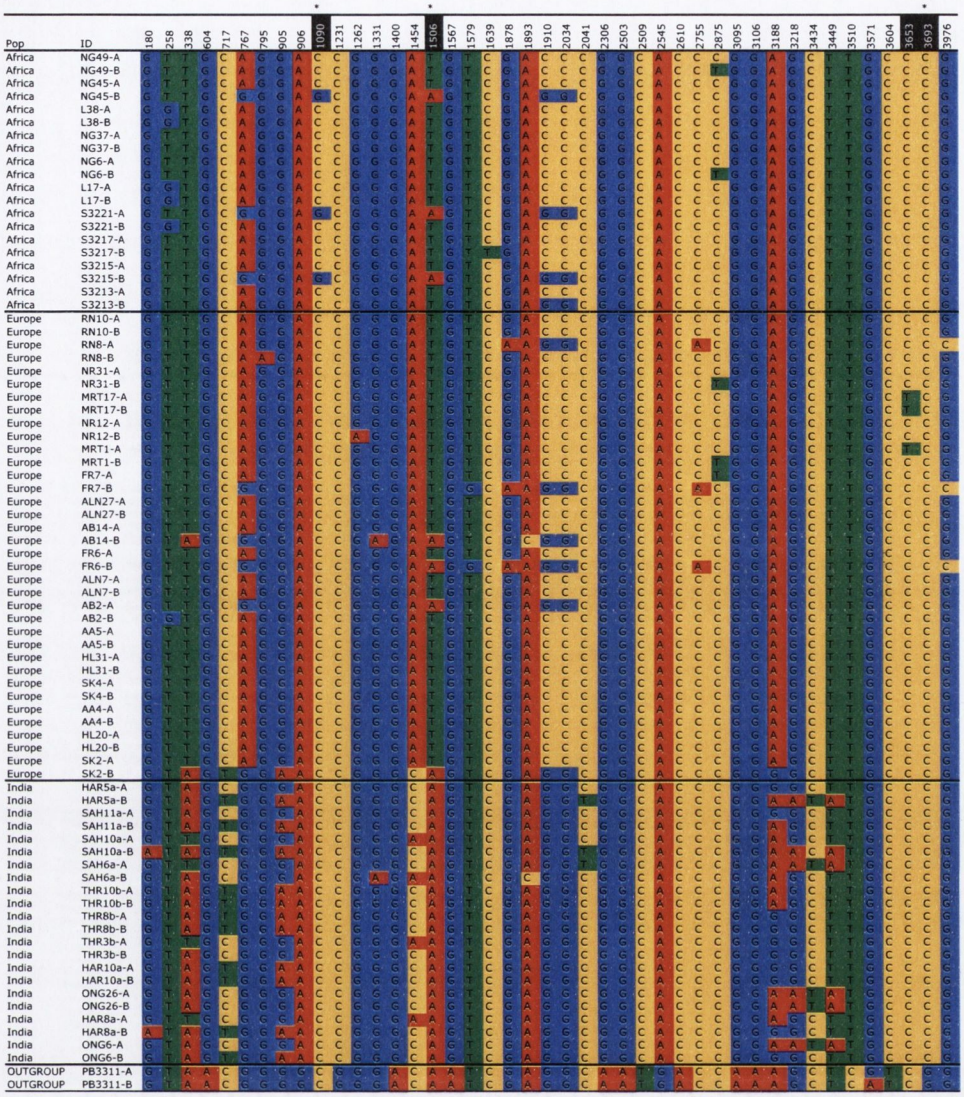
n) TNFAIP8L1

Pop	ID	206	340	901	1541	1554	1589	1622	1670	1679	1714	1803	1805	1883	1990	2089	2183	2308	2341	2342	2391	2586	2632	2732
Africa	NG49-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG49-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG45-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG45-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	T	G	C
Africa	L38-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	L38-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG37-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG37-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG6-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	NG6-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	L17-A	G	G	G	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	L17-B	G	G	G	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3221-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3221-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3217-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3217-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	T	G	C
Africa	S3215-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3215-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3213-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Africa	S3213-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	GB13-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	GB13-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	RN10-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	RN10-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	RN8-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	RN8-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	NR31-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	NR31-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	MRT17-A	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
Europe	MRT17-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
Europe	NR12-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	NR12-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	MRT1-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	MRT1-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	FR7-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	FR7-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	ALN27-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	ALN27-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	AB14-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	AB14-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
Europe	FR6-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	FR6-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	ALN7-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	ALN7-B	G	A	A	C	G	A	T	G	C	G	A	G	C	A	T	T	A	T	T	C	C	A	C
Europe	AB2-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
Europe	AB2-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
Europe	AA5-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	AA5-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	HL31-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	HL31-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	SK4-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	SK4-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
Europe	HL20-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	HL20-B	G	G	A	C	G	A	T	G	C	G	A	G	C	A	T	A	A	C	A	C	C	G	C
Europe	SK2-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
Europe	SK2-B	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
India	HAR5a-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	HAR5a-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	SAH11a-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	SAH11a-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	SAH10a-A	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	SAH10a-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	SAH6a-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	SAH6a-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	THR10b-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	THR10b-B	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	THR8b-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	THR8b-B	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	THR3b-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	THR3b-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	HAR10a-A	G	G	A	C	G	A	T	G	C	G	A	C	C	A	T	A	A	C	A	C	C	G	C
India	HAR10a-B	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	ONG26-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	ONG26-B	G	A	A	C	A	A	T	A	T	G	A	G	T	A	T	A	A	C	A	C	C	G	C
India	HAR8a-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	HAR8a-B	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	ONG6-A	G	A	A	C	G	A	T	G	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
India	ONG6-B	G	A	A	C	A	A	T	A	T	G	A	G	C	A	T	A	A	C	A	C	C	G	C
OUTGROUP	PB3311-A	A	A	A	T	G	C	C	G	C	A	G	G	C	C	C	A	G	C	A	T	C	G	G
OUTGROUP	PB3311-B	A	A	A	T	G	C	C	G	C	A	G	G	C	C	C	A	G	C	A	T	C	G	G

o)TRIF


Pop
ID
Africa NG49-A
Africa NG49-B
Africa NG45-A
Africa NG45-B
Africa L38-A
Africa L38-B
Africa NG37-A
Africa NG37-B
Africa NG6-A
Africa NG6-B
Africa L17-A
Africa L17-B
Africa S3221-A
Africa S3221-B
Africa S3217-A
Africa S3217-B
Africa S3215-A
Africa S3215-B
Africa S3213-A
Africa S3213-B
Europe GB13-A
Europe GB13-B
Europe RN10-A
Europe RN10-B
Europe RN8-A
Europe RN8-B
Europe NR31-A
Europe NR31-B
Europe MRT17-A
Europe MRT17-B
Europe NR12-A
Europe NR12-B
Europe MRT1-A
Europe MRT1-B
Europe FR7-A
Europe FR7-B
Europe ALN27-A
Europe ALN27-B
Europe AB14-A
Europe AB14-B
Europe FR6-A
Europe FR6-B
Europe ALN7-A
Europe ALN7-B
Europe AB2-A
Europe AB2-B
Europe AA5-A
Europe AA5-B
Europe HL31-A
Europe HL31-B
Europe SK4-A
Europe SK4-B
Europe HL20-A
Europe HL20-B
Europe SK2-A
Europe SK2-B
India HAR5a-A
India HAR5a-B
India SAH11a-A
India SAH11a-B
India SAH10a-A
India SAH10a-B
India SAH6a-A
India SAH6a-B
India THR10b-A
India THR10b-B
India THR8b-A
India THR8b-B
India THR3b-A
India THR3b-B
India HAR10a-A
India HAR10a-B
India ONG26-A
India ONG26-B
India HAR8a-A
India HAR8a-B
India ONG6-A
India ONG6-B
OUTGROUP PB3311-A
OUTGROUP PB3311-B

p) TYROBP


Pop
ID
180
258
338
604
717
767
795
905
906
1090
1231
1262
1331
1400
1454
1506
1567
1579
1639
1878
1893
1910
2034
2041
2306
2503
2509
2545
2610
2755
2875
3095
3106
3188
3218
3434
3449
3510
3571
3604
3653
3693
3976
Africa NG49-A
Africa NG49-B
Africa NG45-A
Africa NG45-B
Africa L38-A
Africa L38-B
Africa NG37-A
Africa NG37-B
Africa NG6-A
Africa NG6-B
Africa L17-A
Africa L17-B
Africa S3221-A
Africa S3221-B
Africa S3217-A
Africa S3217-B
Africa S3215-A
Africa S3215-B
Africa S3213-A
Africa S3213-B
Europe RN10-A
Europe RN10-B
Europe RN8-A
Europe RN8-B
Europe NR31-A
Europe NR31-B
Europe MRT17-A
Europe MRT17-B
Europe NR12-A
Europe NR12-B
Europe MRT1-A
Europe MRT1-B
Europe FR7-A
Europe FR7-B
Europe ALN27-A
Europe ALN27-B
Europe AB14-A
Europe AB14-B
Europe FR6-A
Europe FR6-B
Europe ALN7-A
Europe ALN7-B
Europe AB2-A
Europe AB2-B
Europe AA5-A
Europe AA5-B
Europe HL31-A
Europe HL31-B
Europe SK4-A
Europe SK4-B
Europe AA4-A
Europe AA4-B
Europe HL20-A
Europe HL20-B
Europe SK2-A
Europe SK2-B
India HAR5a-A
India HAR5a-B
India SAH11a-A
India SAH11a-B
India SAH10a-A
India SAH10a-B
India SAH6a-A
India SAH6a-B
India THR10b-A
India THR10b-B
India THR8b-A
India THR8b-B
India THR3b-A
India THR3b-B
India HAR10a-A
India HAR10a-B
India ONG26-A
India ONG26-B
India HAR8a-A
India HAR8a-B
India ONG6-A
India ONG6-B
OUTGROUP PB3311-A
OUTGROUP PB3311-B

4.3.3 Haplotypes

Haplotypes were inferred using PHASE v2.1.1 (Stephens and Scheet 2005). Details of the genotypes in each individual at these positions are outlined in the form of alignments of haplotypes inferred by PHASE in Figure 4.4. These figures also indicate the polymorphisms that are non-synonymous (indicated by an asterisk), and whether the polymorphism is in an exon (black), a UTR (grey) or intron/intergenic sequence (white).

4.3.3.1 Networks

Inferred haplotypes were used to draw networks illustrating the diversity at each locus (Figure 4.5). Network diagrams show both the frequency of different haplotypes in the sample, and the relationships between those haplotypes. Each haplotype identified in the sample is represented by a node whose area is proportional to the frequency of that haplotype in the sample. The nodes are joined by branches whose lengths are proportional to the number of mutations separating the nodes. Given the variation in levels of diversity in the sequenced regions, the networks are drawn on different scales. For each gene, the branch length corresponding to one mutation is indicated.

In addition to illustrating the frequency and diversity of haplotypes in the sample, networks can also convey information on the geographical occurrence of haplotypes. In Figure 4.5, *Bos indicus* haplotypes are indicated by yellow nodes, European *Bos taurus* haplotypes are indicated by red nodes and African *Bos taurus* haplotypes are indicated by blue nodes. European and African *Bos taurus* haplotypes are typically closely related, with the majority of haplotypes being separated by only a small number of mutations. In contrast, *Bos indicus* tends to have more low frequency haplotypes separated by greater numbers of mutational events.

A classical signal of expansion can be seen in *Bos taurus* in a number of the networks. The best example is marked § in TYROBP Figure 4.5 . This signal consists of a high frequency central haplotype with a number of lower frequency haplotypes separated by one or two mutations. Other examples, though less clear are marked with § in IL13, TNFAIP8L1, TRIF, ART4 and CD2.

The networks diagrams also highlight a peculiarity of the dataset. In a number of genes one or more European *Bos taurus* haplotypes appears to cluster with *Bos indicus* haplotypes. The networks of CD2, IL2, IL13, TNFAIP8L1, TRIF and TYROBP all

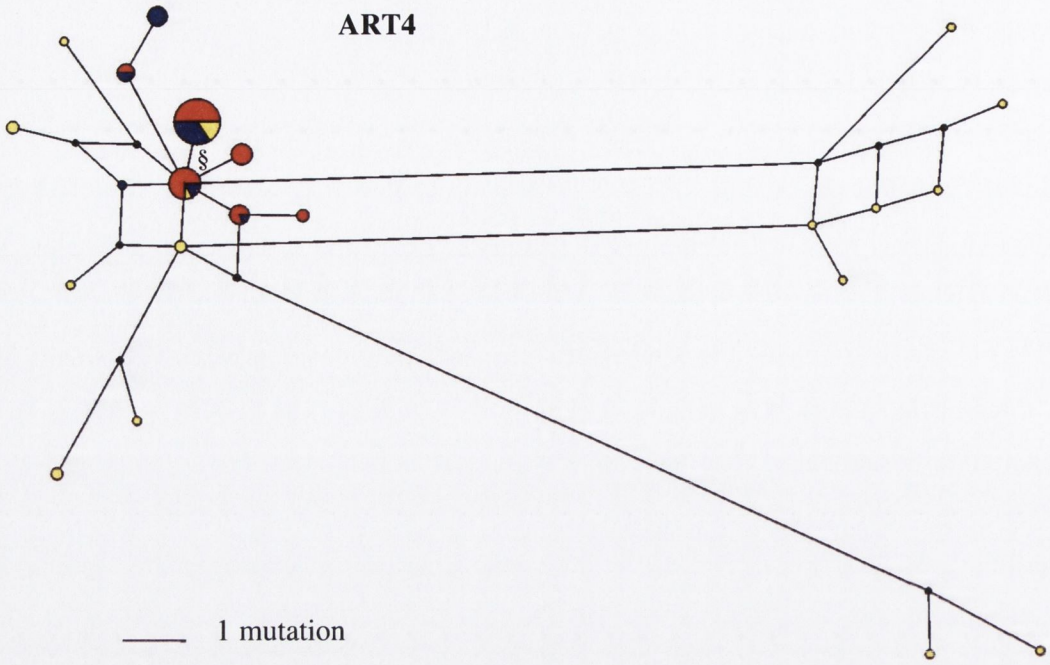
exhibit this feature. Examination of the data indicates that these haplotypes are accounted for by four individuals, Anatolian Black 2 (indicated by * on the networks), Anatolian Black 14 (**), Sikias 2 (***) and Romagnola 8 (****). These breeds are all Mediterranean breeds of cattle: Anatolian Black originated in Turkey, Sikias originated in Greece and Romagnola originated in Italy. These peculiarities are likely to represent *Bos indicus* introgression into the breeds. No such pattern was observed in Western and Northern European *Bos taurus* haplotypes.

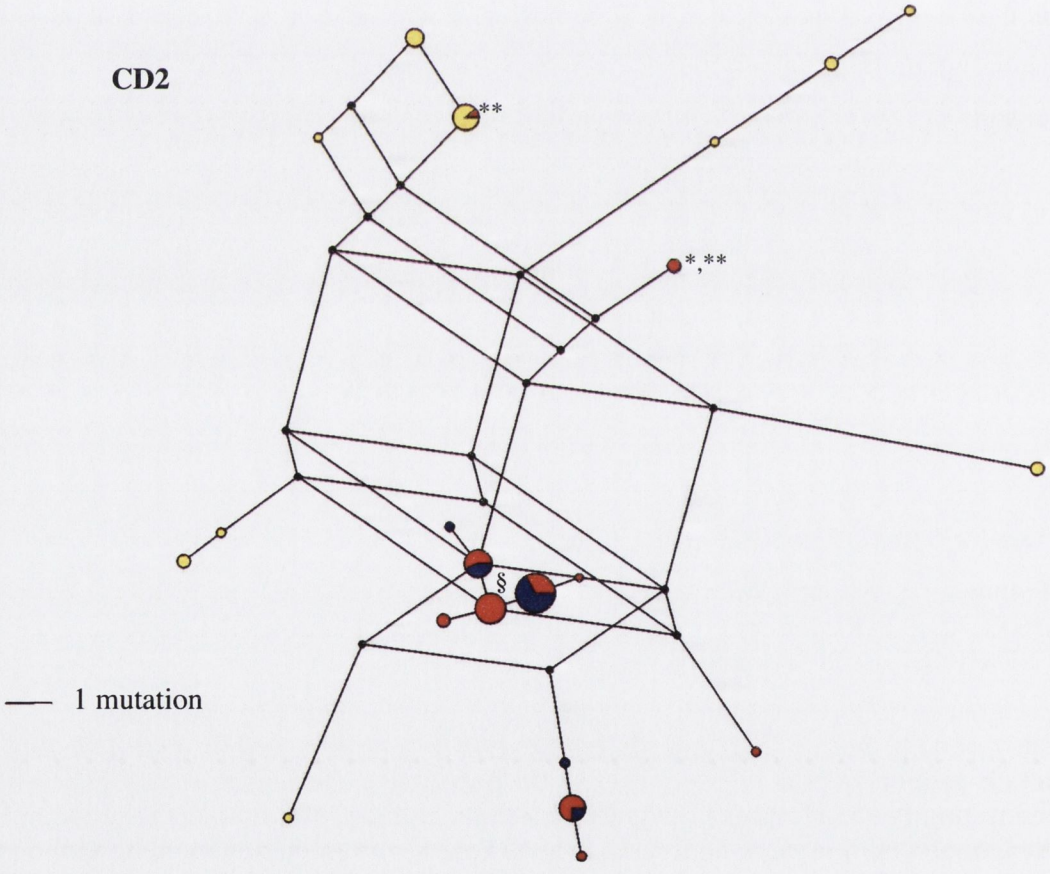
A three nucleotide insertion/deletion in the glycine stretch of the FEZL protein has been reported to be associated with susceptibility to mastitis in Chinese cattle (Sugimoto et al. 2006). The alleles were designated 13G (the allele associated with susceptibility to mastitis) and 12G, referring to the number of glycines in the repeat. In this study, only three animals were identified that carried the 12 G alleles, one of which was heterozygous (homozygous for the deletion: Sikias 2, Friesian 7; and heterozygous for the deletion: Norwegian Red 31). The SK2 haplotypes are part of a high frequency haplotype group but the FR7 haplotypes and the NR31 haplotype, which are identical to each other, form a distinct group of their own. These two haplotypes are indicated with # on the FEZL network.

Figure 4.5 Reduced median (RM) networks.

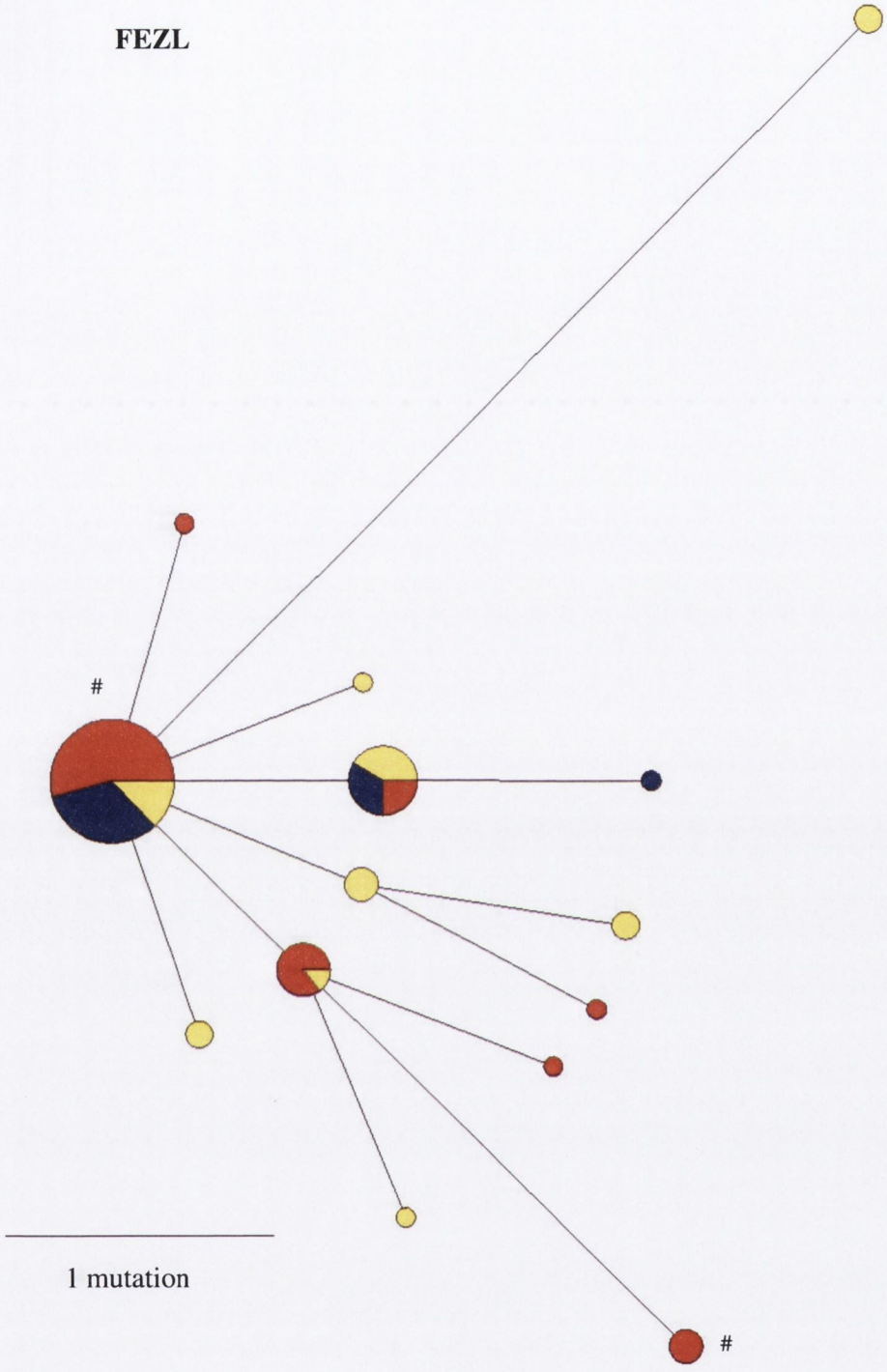
(Following pages.) RM networks for a) ART4, b) CD2, c) FEZL, d) IL2, e) IL13, f) MRPL30, g) MRPS05, h) MRPS14, i)PRF1, j)TLR2, k) TLR4,, l)TLR9 m) TNFAIP8L1, n) TRIF and o) TYROBP.

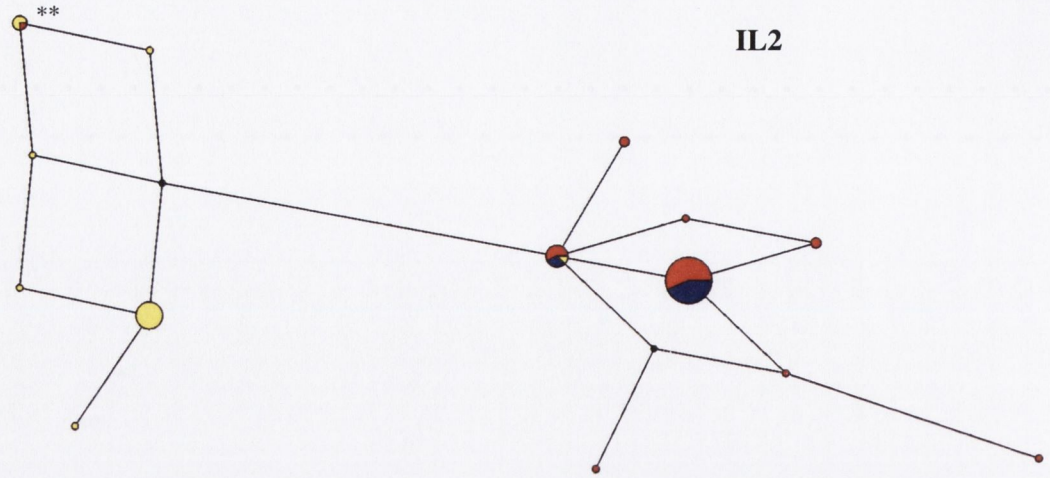
Each network illustrates the haplotypes observed in African *Bos taurus* (blue), European *Bos taurus* (red) and *Bos indicus* (yellow), represented by circles whose area is proportional to the frequency of the haplotype. The length of the branches connecting two haplotypes is proportional to the number of mutations separating the haplotypes. For each figure, the branch length corresponding to one mutation is indicated. Median vectors (haplotypes which are not represented in the sample, but which are inferred to exist or have existed) are indicated by black circles.

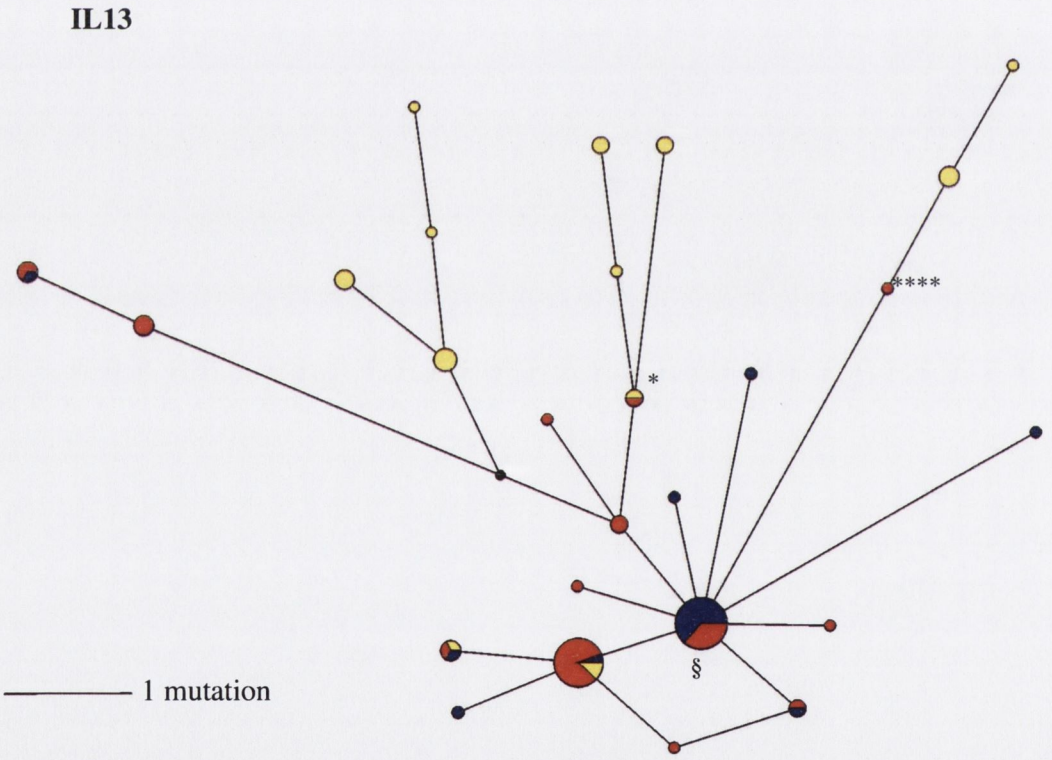

ART4
§
1 mutation

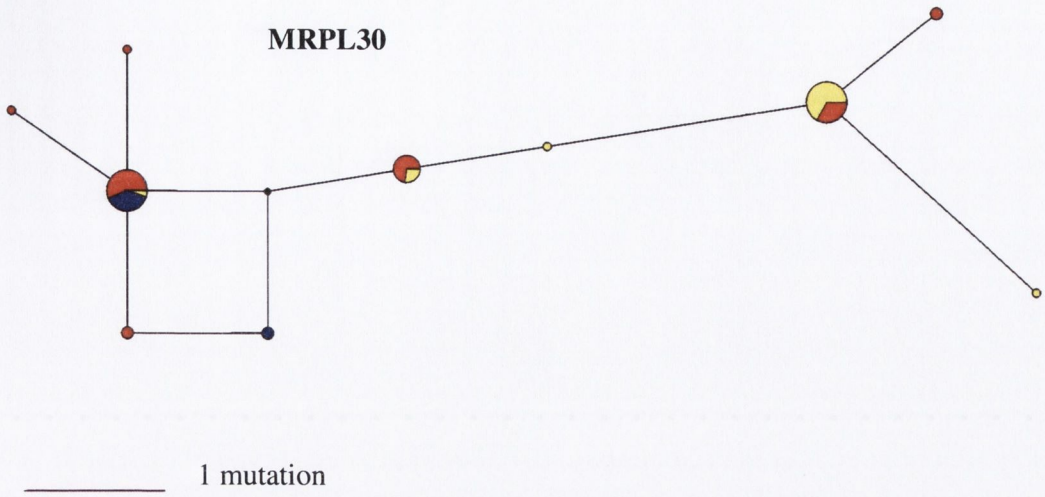

CD2
**
*,**
§
1 mutation

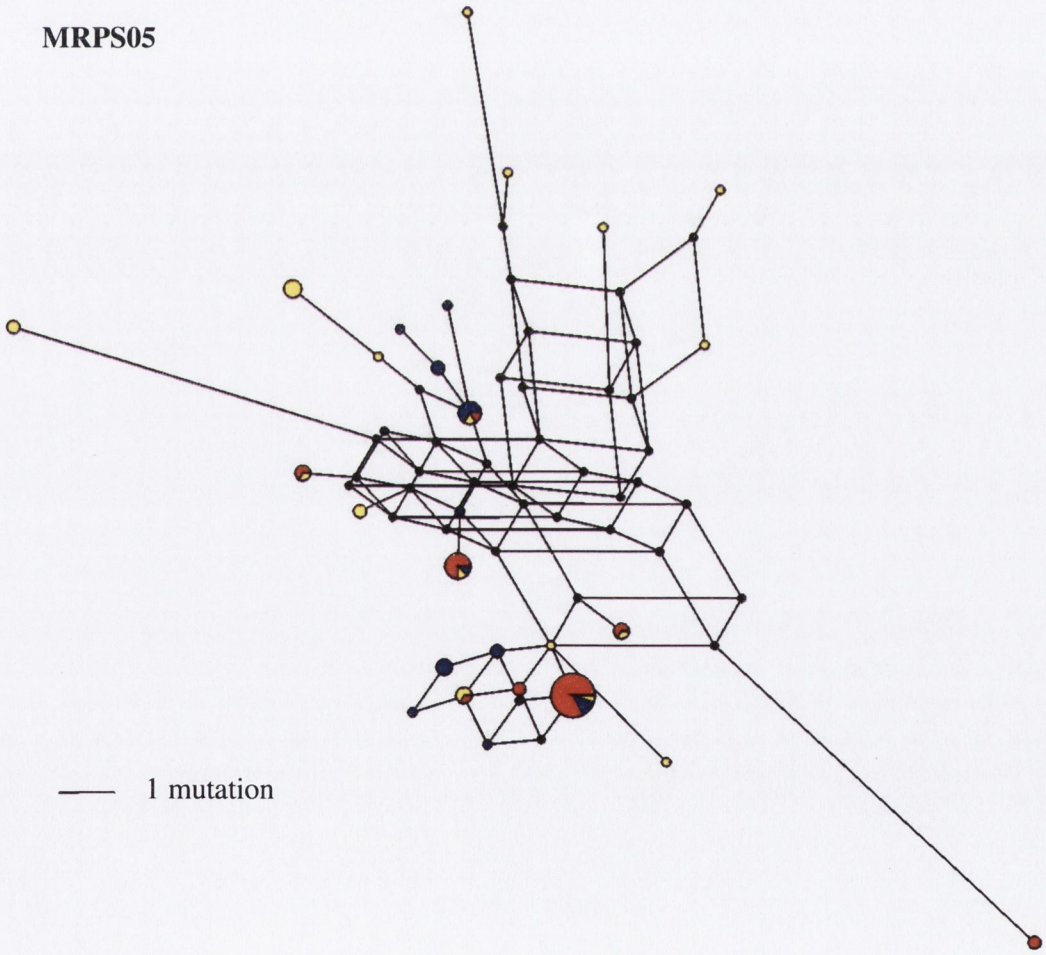
FEZL

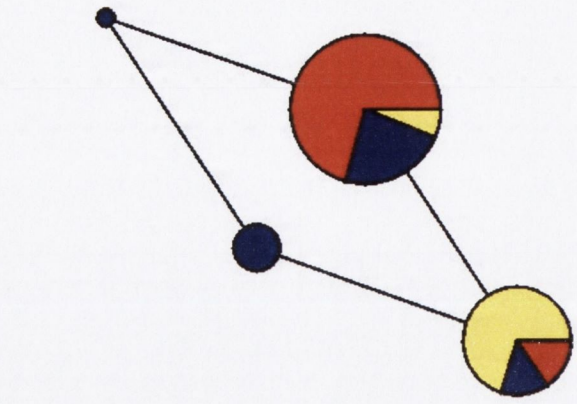



**
IL2
1 mutation

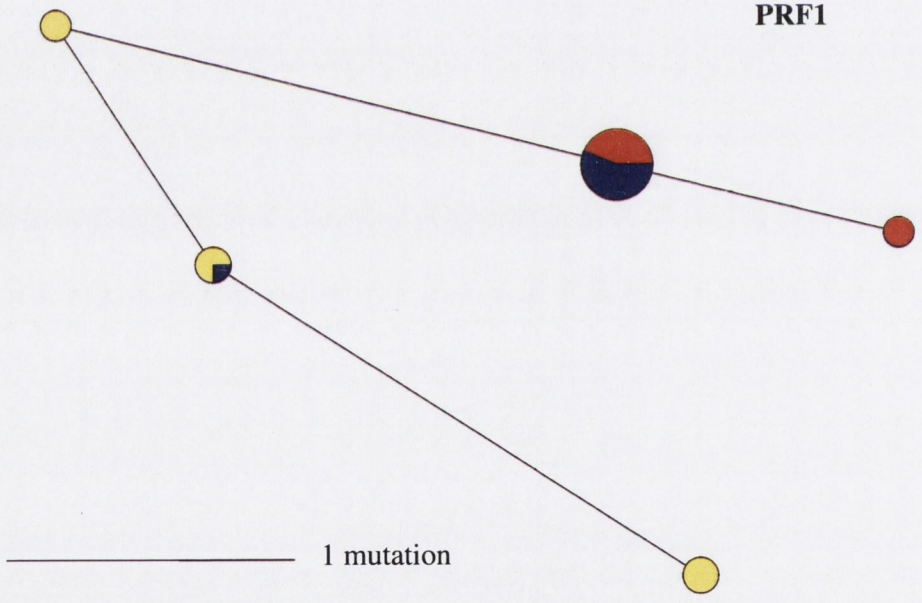

IL13
****
*
§
1 mutation


MRPL30
1 mutation

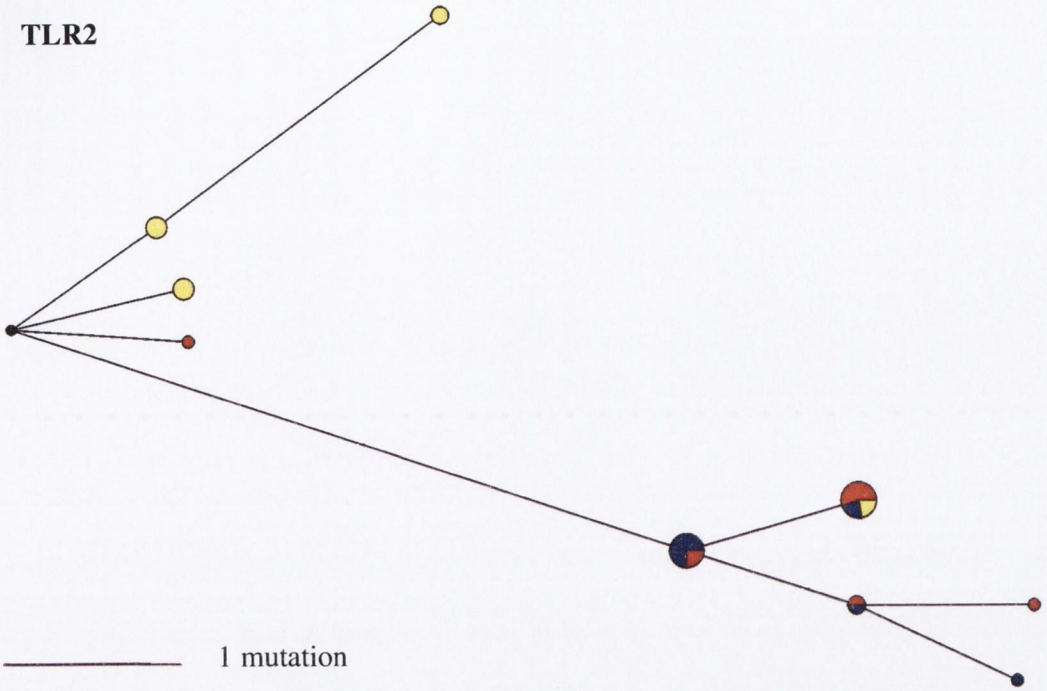

MRPS05
1 mutation

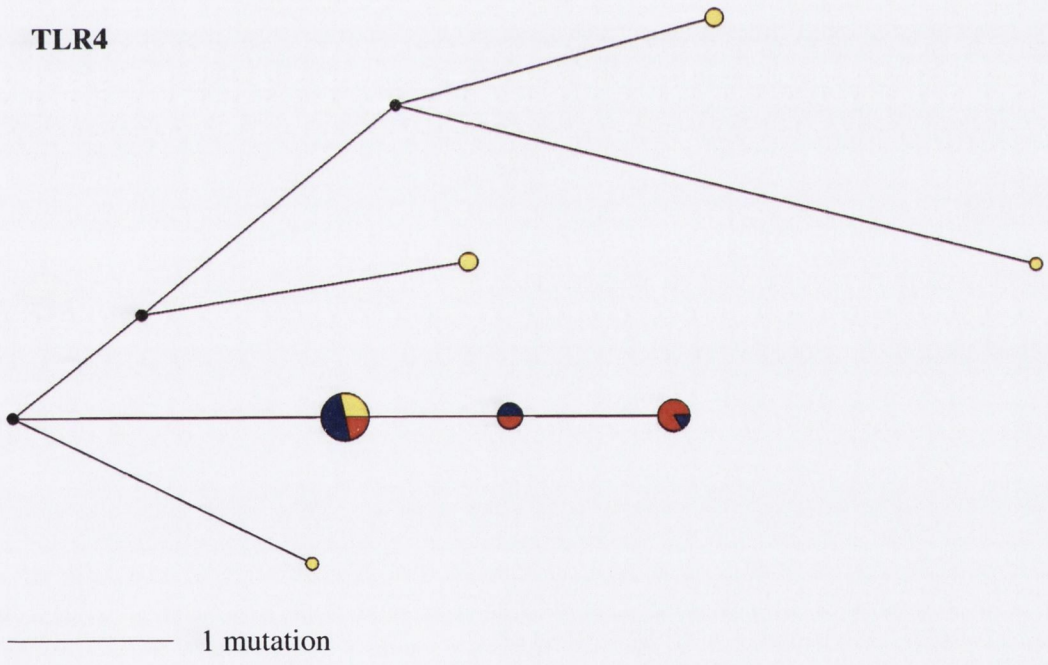


1 mutation

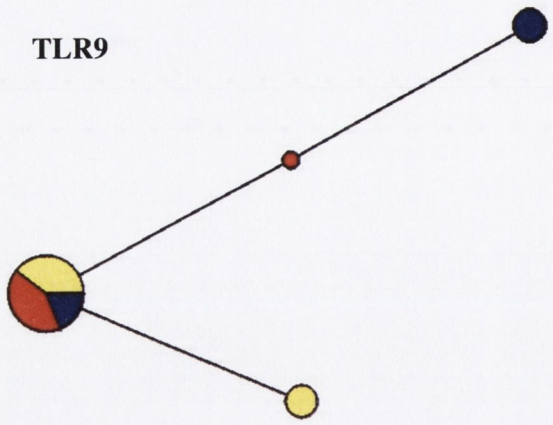


1 mutation


TLR2
1 mutation

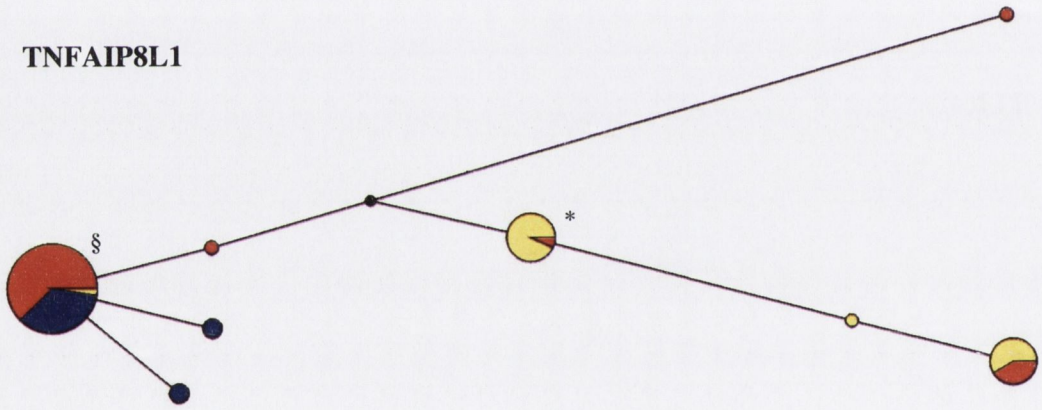

TLR4
1 mutation

TLR9

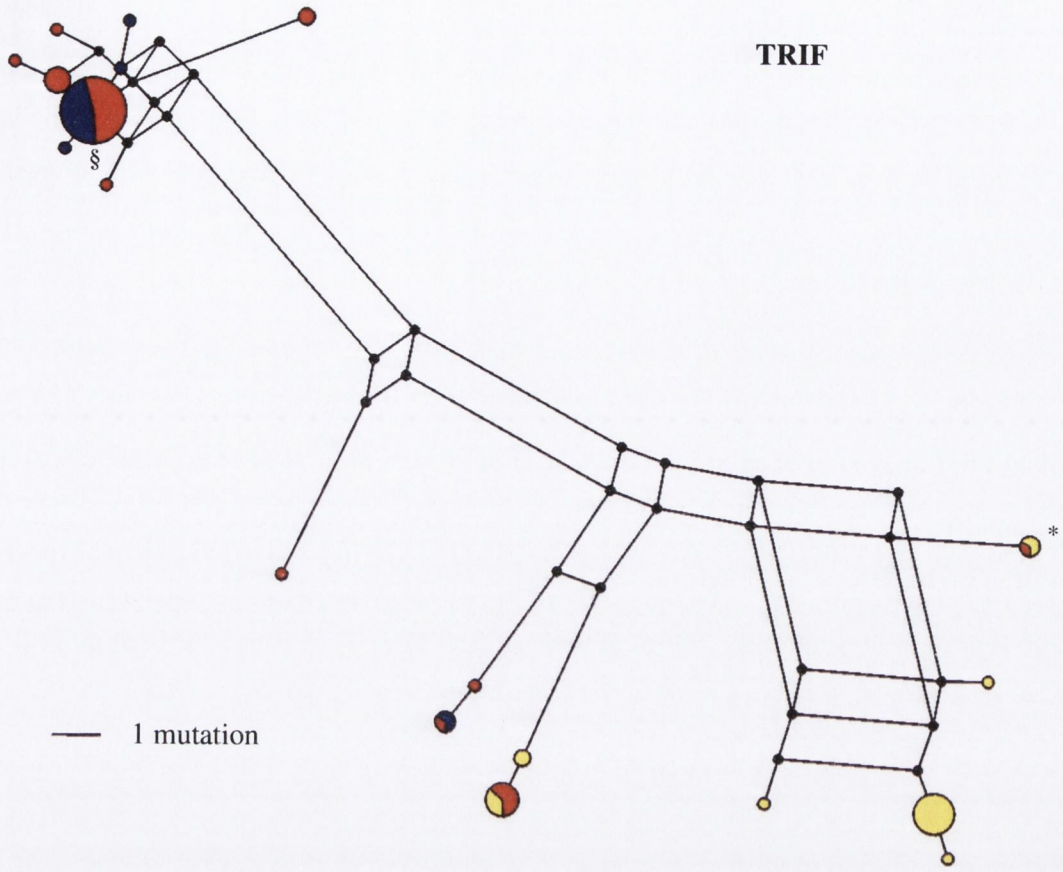


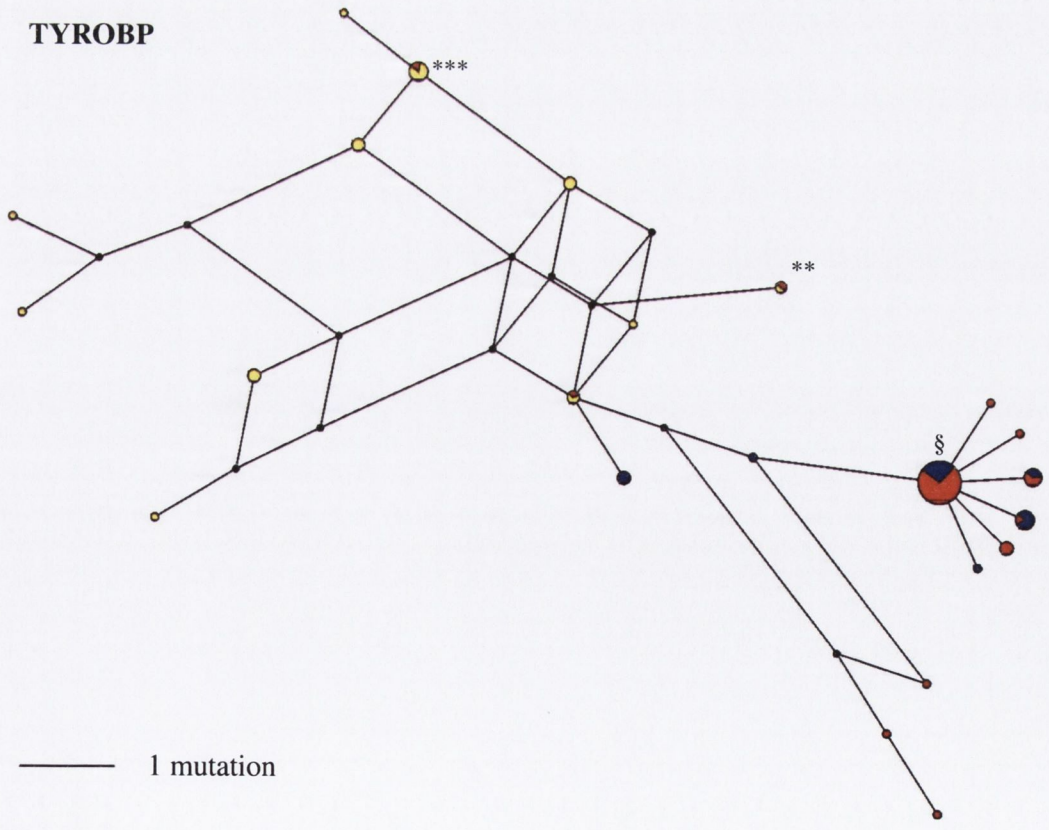
1 mutation

TNFAIP8L1



1 mutation


TRIF
§
*
1 mutation


TYROBP
***
**
§
1 mutation

4.3.3.2 Haplotype Diversity

Haplotype Diversity is a measure of diversity that is analogous to gene diversity (heterozygosity). Haplotype Diversity for each gene in each sample population was calculated in DnaSP (Table 4.4). The significance of the observed haplotype diversity was tested using the *H* test (Depaulis and Veuille 1998) as implemented in DnaSP. The 95% confidence intervals are also indicated.

Overall, haplotype diversity is lowest in African *Bos taurus* (12 out of the 15 genes) and highest in *Bos indicus* (10/15). European *Bos taurus* has the highest value for 3/15 and the lowest for 3/15. Although the African *Bos taurus* typically has the lowest haplotype diversity, there are two genes for which it has the highest haplotype diversity. These are TLR9 and MRPS14. Only one gene, TRIF, shows significantly lower than expected haplotype diversity in all three populations. TRIF also shows unusually high levels of LD (see section 4.3.4)

Table 4.4 Haplotypes and haplotype diversity for each gene in each sample populations (3d.p.). Table continues on following page.

Gene	Population	n[a]	S[b]	No. of haplotypes[c]	Haplotype diversity	Probability[d]	Lower 95% CI[e]	Upper 95% CI[e]
ART4	Africa	20	6	6	0.747	0.599	0.195	0.889
	Europe	36	6	6	0.783	0.878	0.108	0.837
	India	22	33	16	0.961	0.948	0.780	0.970
CD2	Africa	20	13	5	0.616	0.078	0.500	0.921
	Europe	36	30	10	0.846	0.325	0.705	0.941
	India	22	29	10	0.87	0.245	0.732	0.961
FEZL	Africa	18	2	3	0.451	0.561	0.000	0.758
	Europe	36	8	7	0.633	0.636	0.056	0.820
	India	22	11	9	0.887	**0.982**	**0.247**	**0.879**
IL2	Africa	20	1	2	0.268	0.597	0.000	0.679
	Europe	36	12	9	0.614	0.521	0.056	0.833
	India	22	7	7	0.593	0.411	0.091	0.853
IL13	Africa	20	15	9	0.653	0.419	0.190	0.874
	Europe	36	15	13	0.814	0.744	0.300	0.887
	India	22	14	12	0.935	**0.975**	0.580	0.939
MRPL30	Africa	12	2	2	0.303	0.349	0.000	0.788
	Europe	36	9	7	0.763	0.499	0.367	0.890
	India	22	7	5	0.468	0.227	0.091	0.848
MRPS05	Africa	18	11	10	0.915	0.917	0.549	0.935
	Europe	38	19	10	0.728	0.171	0.552	0.922
	India	22	31	16	0.961	0.946	0.797	0.970
MRPS14	Africa	22	2	4	0.645	0.705	0.000	0.797
	Europe	40	1	2	0.185	0.619	0.000	0.621
	India	22	1	2	0.247	0.597	0.000	0.647
PRF1	Africa	10	3	2	0.200	0.374	0.000	0.800
	Europe	10	1	2	0.467	0.691	0.000	0.778
	India	10	3	3	0.733	0.646	0.000	0.889

Gene	Population	n[a]	S[b]	No. of haplotypes[c]	Haplotype diversity	Probability[d]	Lower 95% CI[e]	Upper 95% CI[e]
TLR2	Africa	10	3	4	0.644	0.652	0.000	0.844
	Europe	10	8	5	0.756	0.575	0.200	0.911
	India	10	9	4	0.822	0.482	0.378	0.956
TLR4	Africa	10	2	3	0.511	0.530	0.000	0.822
	Europe	10	2	3	0.689	0.761	0.000	0.844
	India	10	15	5	0.822	0.303	0.533	0.978
TLR5	Africa	10	0	1				
	Europe	10	0	1				
	India	10	1	2	0.533	0.686	0.000	0.778
TLR9	Africa	6	2	2	0.600	0.635	0.000	0.933
	Europe	8	1	2	0.250	0.699	0.000	0.714
	India	10	1	2	0.467	0.658	0.000	0.778
TNFAIP8L1	Africa	20	2	3	0.358	0.621	0.000	0.700
	Europe	36	10	5	0.384	0.085	0.208	0.865
	India	22	6	4	0.571	0.304	0.173	0.870
TRIF	Africa	20	24	5	0.442	**0.006**	**0.568**	**0.937**
	Europe	36	41	11	0.751	**0.013**	**0.803**	**0.963**
	India	22	22	7	0.693	**0.017**	**0.714**	**0.961**
TYROBP	Africa	20	8	6	0.779	0.670	0.268	0.884
	Europe	36	20	12	0.657	0.255	0.340	0.890
	India	22	12	11	0.913	0.957	0.528	0.922

Significant values are indicated in red. a) the number of chromosomes (i.e. twice the number of individuals) in the sample population; b) the number of segregating sites in the sample population; c) the number of haplotypes observed in the population; d) the probability of observing a lower value; e) the 95% confidence intervals as determined by coalescent simulation using DnaSP.

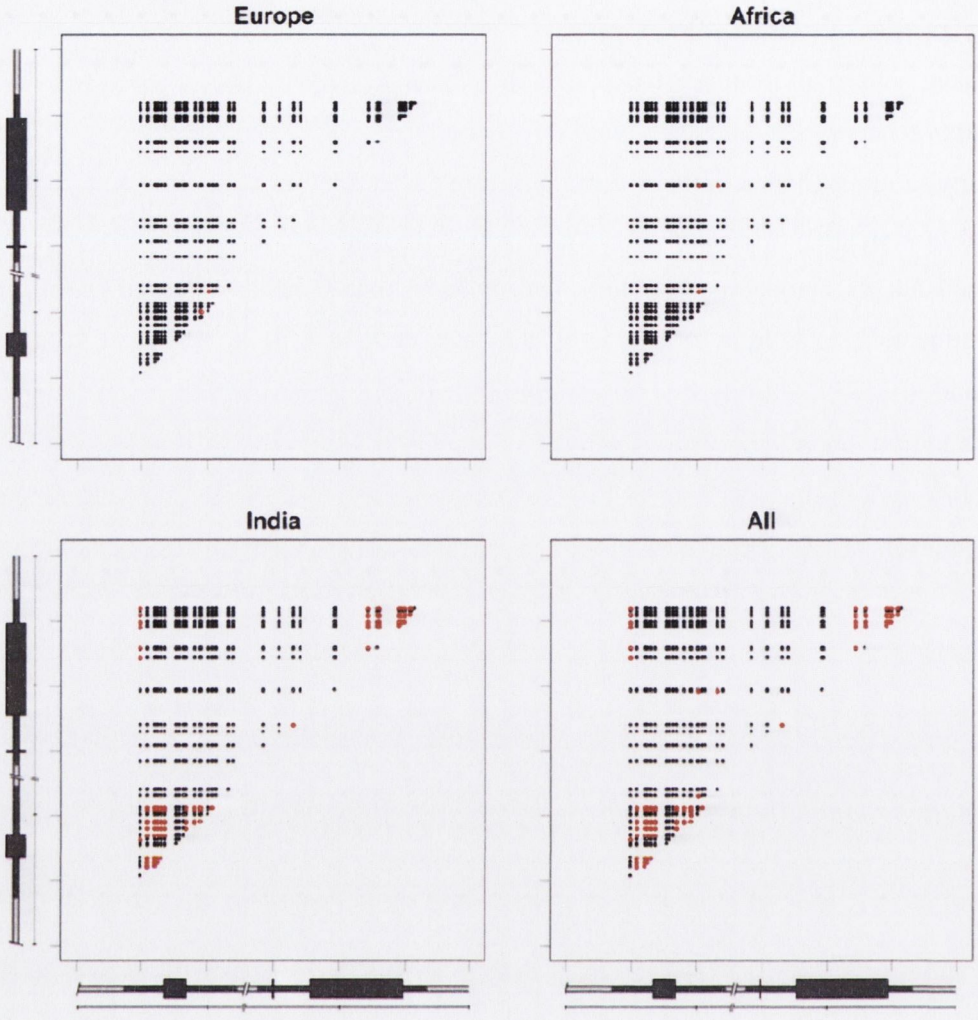
4.3.4 Linkage Diseqilibrium

Linkage Disequilibrium across each of the sequenced regions was tested using GenePop (http://genepop.curtin.edu.au/) (Raymond and Rousset 1995). Figure 4.6 illustrates pairwise comparisons that showed significant LD at the 5% level for each of the genes that had significant LD (red circles) while non-significant comparisons, and sites that could not be compared were indicated by black dots. LD was tested in each of the continental populations, as well as in the combined population. No significant LD was found in MRPS14 (only two polymorphic sites), TLR2, TLR4, TLR5, TLR9 and PRF1. In addition, not all populations had significant LD at all regions. LD tended to be higher in *Bos indicus* than in either of the *Bos taurus* populations. TRIF showed exceptionally high levels of LD, particularly in the European and African *Bos taurus* populations.

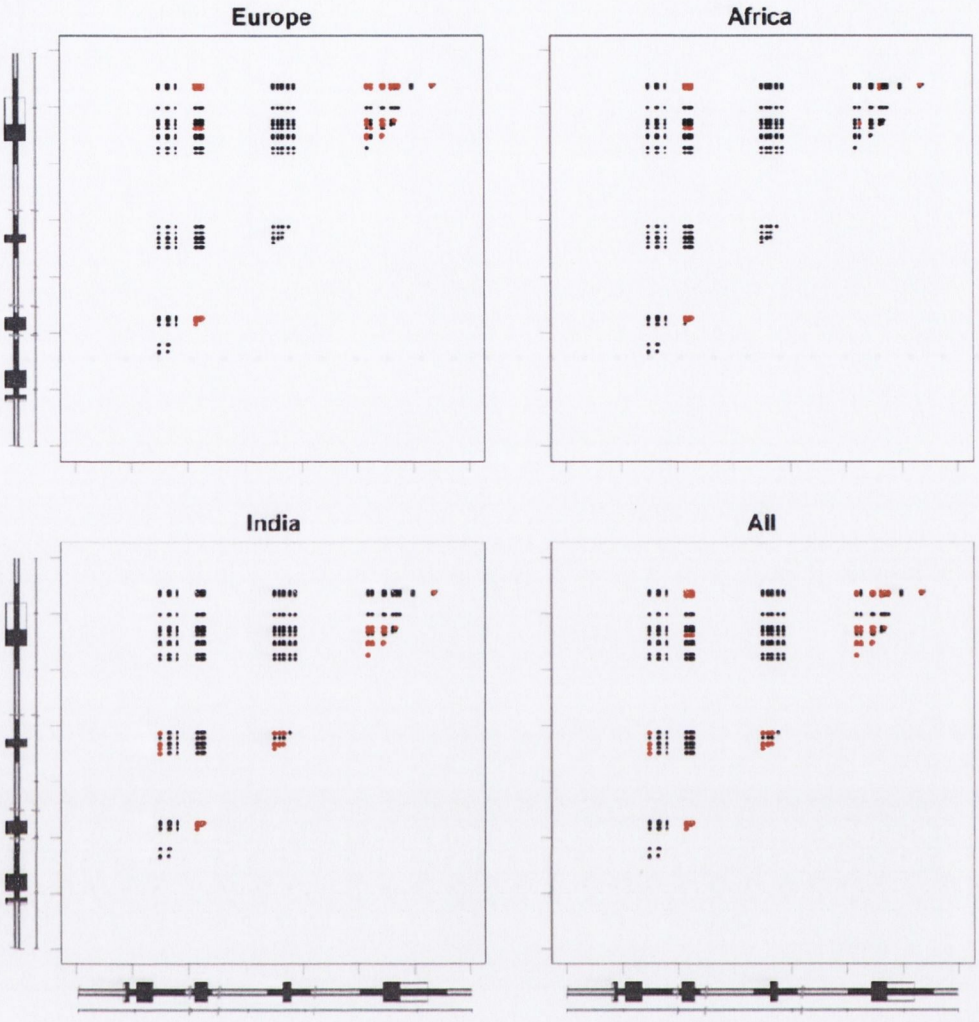
Figure 4.6 Significant pairwise linkage disequilibrium (LD).

(Following pages.) For each gene, pairs of sites with significant linkage disequilibrium at the $p < 0.05$ level are indicated by red dots while all other pairs of sites are indicated by black dots. Significant pairwise LD in European *Bos taurus* (top left) African *Bos taurus* (top right), *Bos indicus* (bottom left) and the combined sample of all populations (bottom right). The structure of each gene is indicated at the left and at the bottom of the LD plots.

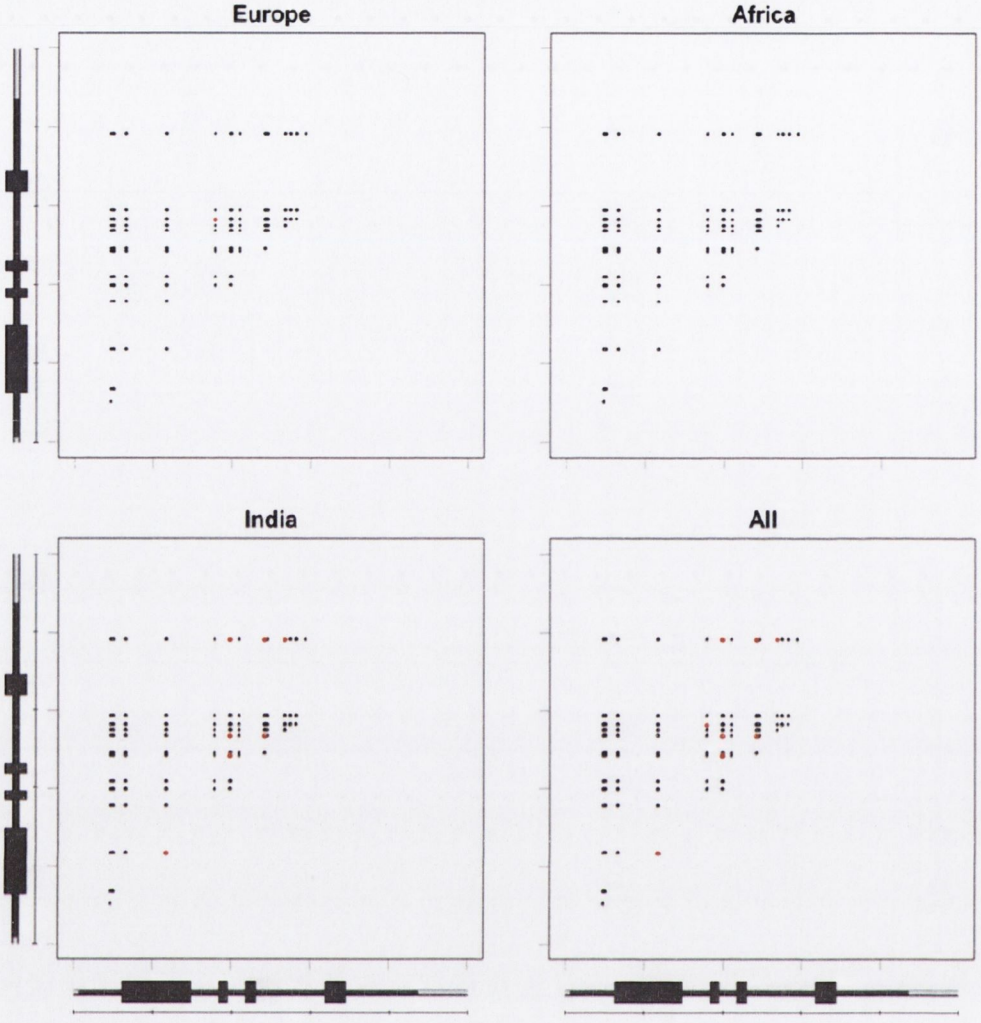
a) ART4



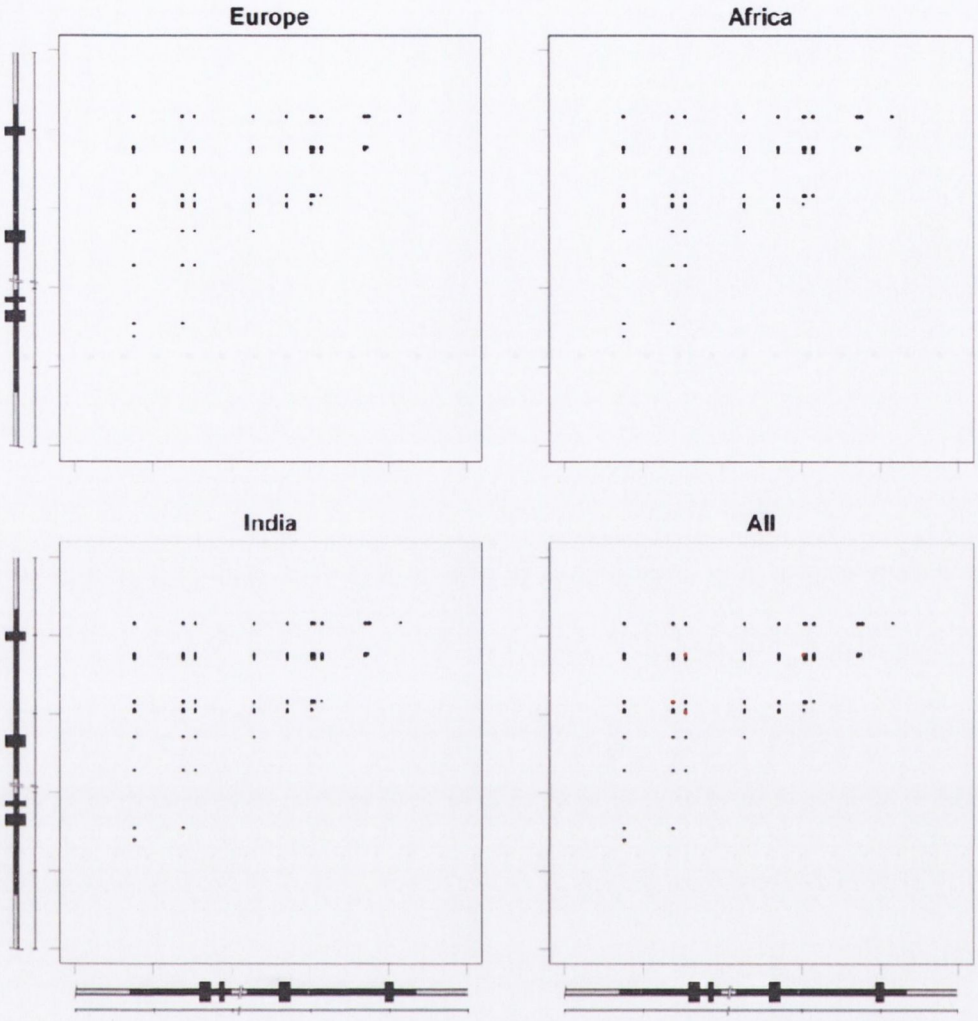
b) CD2



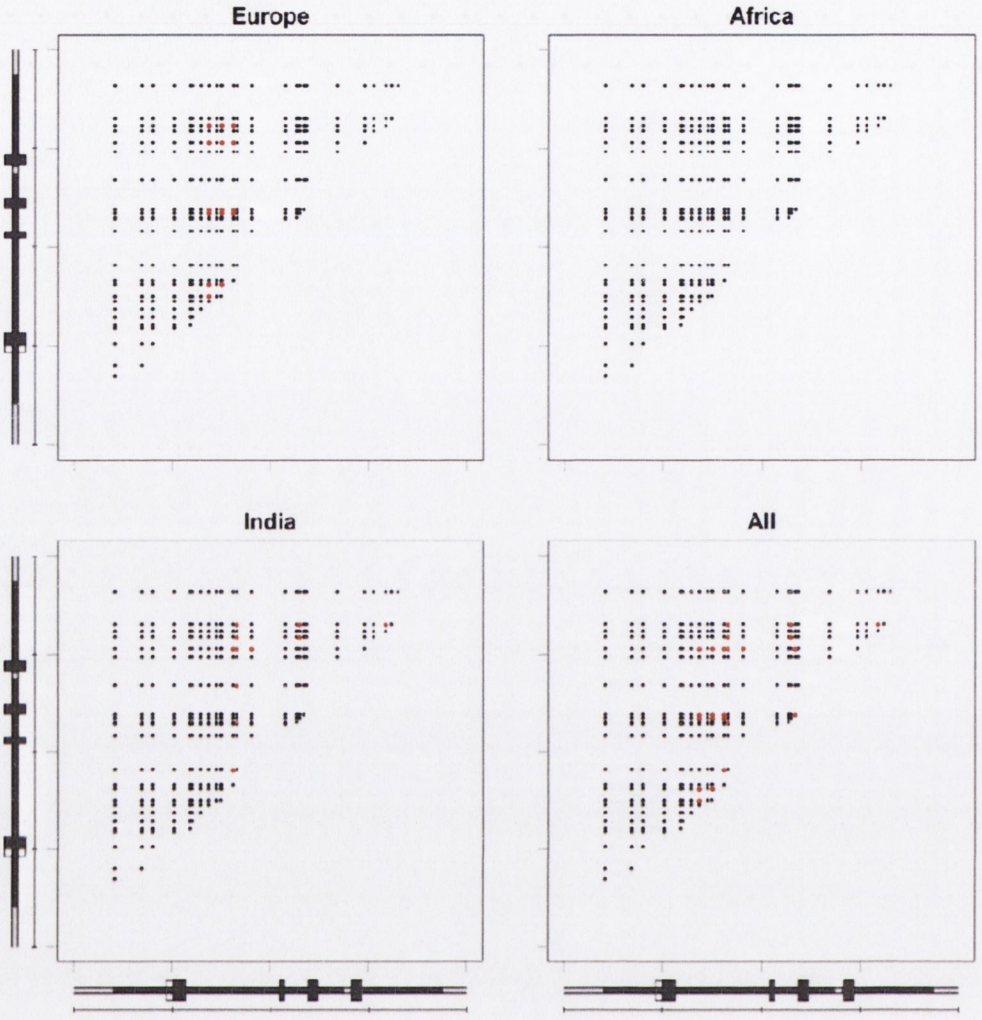
c) FEZL



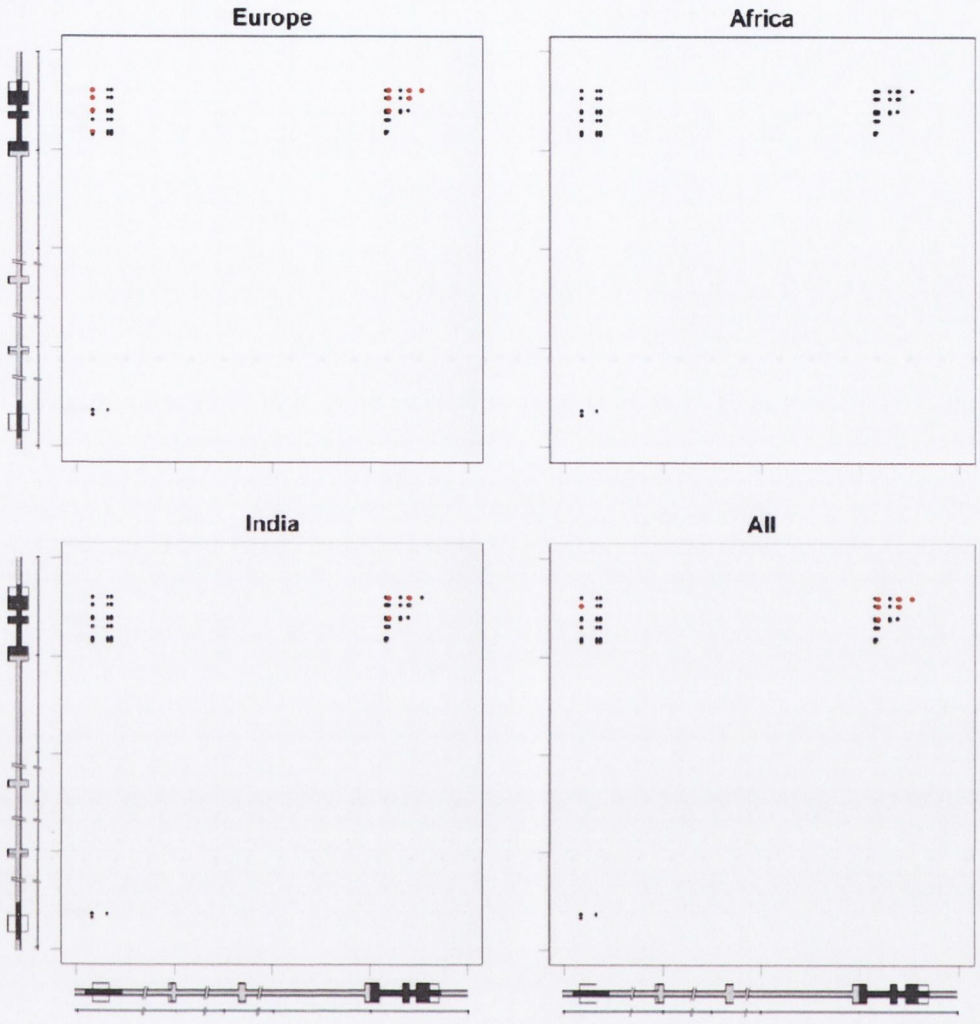
d) IL2



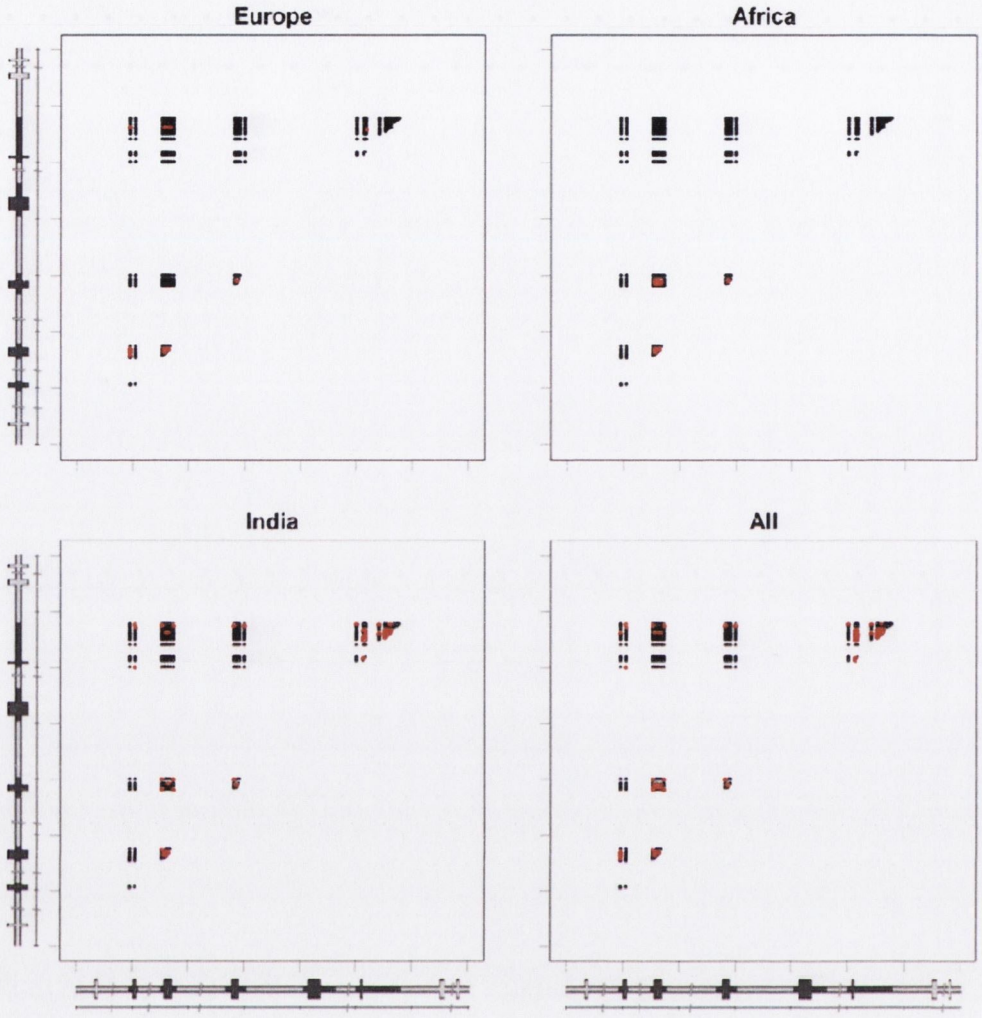
e) IL13



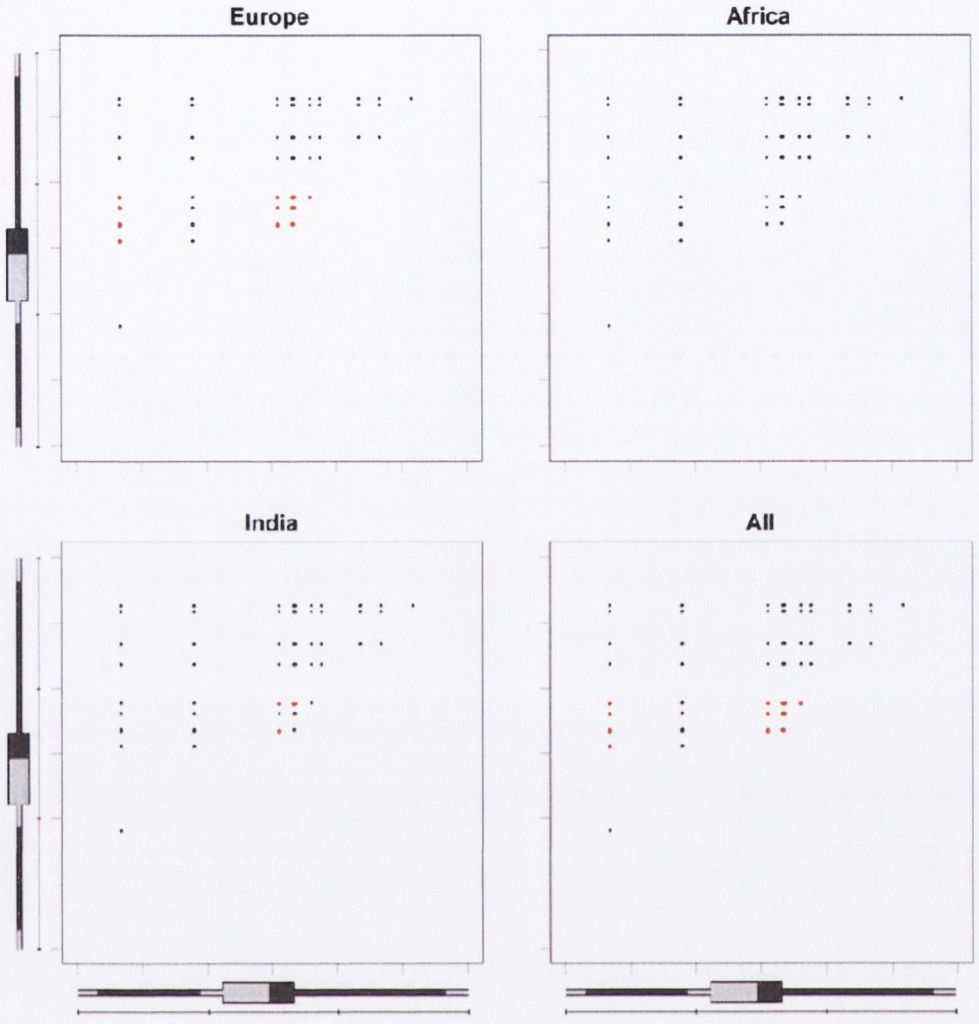
f) MRPL30



g) MRPS05



h) TNFAIP8L1



i) TRIF

j) TYROBP

4.3.5 Neutrality Tests

4.3.5.1 Tajima's *D*

Tajima's *D* compares two estimators of θ: the average number of pairwise differences, π, and the number of segregating sites, *S*, corrected for the sample size, *n*, S/a_n or θ_w (Tajima 1989). A negative value of Tajima's *D* indicates an excess of low frequency variants. Low frequency variants would be expected to be in excess if positive selection had occurred, but might also be observed as a result of certain demographic scenarios, such as population expansion. On the other hand, a positive value of Tajima's *D* indicates an excess of intermediate frequency variants. Such an observation could be a result of balancing selection acting to maintain variation at a particular locus, or could also be observed in a sample with strong population structure.

Tajima's *D* was calculated using DnaSP (Rozas et al. 2003). Significance levels and 95% confidence intervals were determined by coalescent simulation under a neutral model in DnaSP. The values of Tajima's *D*, as calculated using DnaSP, are given in Table 4.5. Several significantly negative values for Tajima's *D* were observed: IL2 in European *Bos taurus* ($p=0.015$); IL13 in African *Bos taurus* ($p<0.001$) and PRF1 in African *Bos taurus* ($p=0.023$). In addition, TLR9 in African *Bos taurus* was found to have a significantly positive value of Tajima's *D* ($p=0.984$ which corresponds to $p=0.016$ of observing a higher value).

Table 4.5 Nucleotide diversity and Tajima's *D* values obtained for each gene in each sample population. Table continues on following page.

Gene	Population	n[a]	S[b]	θ (per site)[c]	π (per site)[d]	Tajima's *D*	Probability[e]	Tajima's *D* 95% CI Lower[f]	Tajima's *D* 95% CI Upper[f]
ART4	Africa	20	6	0.00078	0.00096	0.721	0.784	-1.638	1.975
	Europe	36	6	0.00067	0.00064	-0.146	0.512	-1.629	2.086
	India	22	33	0.00420	0.00456	0.332	0.706	-1.758	1.774
CD2	Africa	20	13	0.00098	0.00099	0.051	0.556	-1.714	1.721
	Europe	36	30	0.00193	0.00161	-0.576	0.295	-1.616	1.917
	India	22	29	0.00212	0.00225	0.236	0.653	-1.740	1.767
FEZL	Africa	18	2	0.00014	0.00013	-0.192	0.371	-1.508	1.784
	Europe	36	8	0.00045	0.00026	-1.244	0.124	-1.724	1.976
	India	22	11	0.00070	0.00050	-1.018	0.184	-1.779	1.804
IL2	Africa	20	1	0.00008	0.00008	-0.086	0.329	-1.513	1.639
	Europe	36	12	0.00084	0.00037	-1.754	**0.015**	**-1.667**	**1.973**
	India	22	7	0.00055	0.00042	-0.769	0.261	-1.642	1.794
IL13	Africa	20	15	0.00128	0.00053	-2.156	**< 0.001**	**-1.752**	**1.936**
	Europe	36	15	0.00110	0.00076	-0.998	0.166	-1.700	1.785
	India	22	14	0.00116	0.00141	0.741	0.800	-1.729	1.893
MRPL30	Africa	12	2	0.00054	0.00050	-0.248	0.350	-1.451	1.753
	Europe	36	9	0.00178	0.00213	0.582	0.638	-1.755	1.936
	India	22	7	0.00157	0.00116	-0.848	0.222	-1.656	1.906
MRPS05	Africa	18	11	0.00130	0.00179	1.363	0.924	-1.849	1.807
	Europe	38	19	0.00184	0.00164	-0.362	0.396	-1.734	1.830
	India	22	31	0.00346	0.00403	0.628	0.791	-1.789	1.733
MRPS14	Africa	22	2	0.00087	0.00147	1.542	0.938	-1.534	1.857
	Europe	40	1	0.00037	0.00029	-0.307	0.278	-1.276	1.656
	India	22	1	0.00044	0.00039	-0.175	0.296	-1.240	1.581
PRF1	Africa	10	3	0.00096	0.00054	-1.562	**0.023**	**-1.401**	**1.641**
	Europe	10	1	0.00032	0.00042	0.820	0.825	-1.401	1.641
	India	10	3	0.00096	0.00139	1.604	0.962	-1.667	1.772

Gene	Population	n[a]	S[b]	θ (per site)[c]	π (per site)[d]	Tajima's *D*	Probability[e]	Tajima's *D* 95% CI Lower[f]	Tajima's *D* 95% CI Upper[f]
TLR2	Africa	10	3	0.00095	0.00083	-0.5067	0.332	-1.562	1.830
	Europe	10	8	0.00254	0.00192	-1.093	0.176	-1.687	1.671
	India	10	9	0.00286	0.00319	0.454	0.672	-1.667	1.788
TLR4	Africa	10	2	0.00064	0.00061	-0.184	0.319	-1.562	1.641
	Europe	10	2	0.00064	0.00093	1.439	0.932	-1.562	1.844
	India	10	15	0.00482	0.00471	-0.108	0.484	-1.796	1.674
TLR5	Africa	10	0	0.00000	0.00000				
	Europe	10	0	0.00000	0.00000				
	India	10	1	0.00086	0.00129	1.303	0.931	-1.562	1.641
TLR9	Africa	6	2	0.00093	0.00128	1.753	**0.984**	**-1.295**	**1.648**
	Europe	8	1	0.00041	0.00027	-1.055	0.192	-1.310	1.444
	India	10	1	0.00038	0.00050	0.830	0.839	-1.562	1.641
TNFAIP8L1	Africa	20	2	0.00026	0.00018	-0.769	0.198	-1.513	1.683
	Europe	36	10	0.00111	0.00089	-0.616	0.303	-1.685	1.987
	India	22	6	0.00076	0.00078	0.096	0.602	-1.762	1.831
TRIF	Africa	20	24	0.00164	0.00109	-1.283	0.089	-1.663	1.742
	Europe	36	41	0.00239	0.00244	0.073	0.564	-1.688	1.875
	India	22	22	0.00146	0.00186	1.027	0.864	-1.669	1.859
TYROBP	Africa	20	8	0.00057	0.00055	-0.137	0.482	-1.638	1.869
	Europe	36	20	0.00122	0.00064	-1.607	0.036	-1.713	1.926
	India	22	12	0.00083	0.00101	0.745	0.798	-1.761	1.861

Significant values are indicated in red. a) the number of chromosomes (i.e. twice the number of individuals) in the sample population; b) the number of segregating sites in the sample population; c) the number of segregating sites, adjusted for sample size, θ_w (per site); d) the average number of pairwise differences, π (per site) e) the probability of observing a lower value; f) the 95% confidence intervals as determined by coalescent simulation using DnaSP.

4.3.5.2 Fu and Li's tests

Fu and Li's tests (Fu and Li 1993) were carried out both with (*D* and *F*) and without (*D** and *F**) an outgroup, with the exception of those genes that did not have an outgroup sequence. Fu and Li favour the use of the *D* and *F* tests over the *D** and *F** wherever possible. Given that it was not possible to carry out *D* and *F* tests for all genes, *D** and *F** were carried out for all genes in order that reasonable comparisons could be made across all genes. Fu and Li's tests are based on the idea that if directional selection has been acting on a region, there will be an excess of recent mutations, i.e., mutations that lie on the external branches of the genealogy. Conversely if balancing selection has been acting on a region, there will be an excess of older mutations, mutations which lie on the internal branches of the genealogy. A negative value of indicates that there is an excess of mutations on the external branches, while a positive value indicates that there is a deficiency.

Fu and Li advocate the use of the *D* and *F* statistics, which require an outgroup sequence to determine whether mutations are internal of external on the genealogy, over the *D** and *F** statistics, which are based on the number of singletons (Fu and Li 1993). Fu and Li's *D* and *F* were significantly negative in IL13 in African *Bos taurus*, and significantly positive in ART4 in *Bos indicus* (Table 4.6). No other genes/population analyses showed a significant result for Fu and Li's F, but several showed significant values of Fu and Li's *D*. IL13 in *Bos indicus* showed a significantly positive value of *D*, while IL13 in Europe was negative and marginally significant. PRF1 (African *Bos taurus*), TLR2 (*Bos indicus*), and TRIF (African *Bos taurus*) had marginally significant positive values.

Fu and Li's *D** and *F** were also calculated (Table 4.7). Similar results to those obtained for Fu and Li's *D* and *F* were obtained with Fu and Li's *D** and *F**. IL2 in European *Bos taurus* had a significantly negative value of *D** and *F**. A notable exception is PRF1 in African *Bos taurus*, which had a positive and marginally significant value of *D*, but a significantly negative value of *D**. The cause of this anomaly is apparent in Figure 4.4(i). Fu and Li's D* relies on the genetic variation within the population to predict whether the mutation is ancestral or derived, while Fu and Li's D uses an outgroup sequence. In this population, the minor allele at all three polymorphis sites is present in the outgroup sequence (Gaur) and therefore the two statistics result in different estimates of ancestral or derived status.

Table 4.6 Fu and Li's *D* and *F* (tests with an outgroup) values obtained for each gene in each sample population. Table continues on following page.

Gene	Population	n[a]	S[b]	Fu & Li's *D*	Probability[c]	Fu & Li's *D* 95% CI Lower[d]	Fu & Li's *D* 95% CI Upper[d]	Fu & Li's *F*	Probability[c]	Fu & Li's *F* 95% CI Lower[d]	Fu & Li's *F* 95% CI Upper[d]
ART4	Africa	20	6	-0.239	0.405	-2.385	1.487	0.048	0.519	-2.330	1.637
	Europe	36	6	1.234	0.937	-2.570	1.432	0.953	0.809	-2.444	1.754
	India	22	33	1.747	**0.993**	**-2.197**	**1.488**	1.549	0.967	-2.352	1.674
CD2	Africa	20	13	1.193	0.911	-2.389	1.573	1.008	0.834	-2.519	1.832
	Europe	36	30	0.611	0.720	-2.151	1.576	0.221	0.623	-2.372	1.772
	India	22	29	1.200	0.921	-2.197	1.605	1.068	0.876	-2.425	1.743
FEZL	Africa	18	2	-0.627	0.281	-2.125	1.150	-0.597	0.261	-2.288	1.425
	Europe	36	8	-1.475	0.119	-2.449	1.310	-1.666	0.089	-2.394	1.683
	India	22	11	0.528	0.667	-2.190	1.477	0.056	0.511	-2.192	1.838
IL2	Africa	20	1	0.627	0.624	-1.596	1.012	0.512	0.718	-2.330	1.243
	Europe	36	12	-1.144	0.133	-2.220	1.310	-1.613	0.095	-2.359	1.717
	India	22	7	-1.479	0.109	-2.224	1.425	-1.507	0.114	-2.446	1.674
IL13	Africa	20	15	-3.521	**< 0.001**	**-2.114**	**1.436**	-3.731	**< 0.001**	**-2.330**	**1.683**
	Europe	36	15	-2.034	0.034	-2.207	1.503	-2.020	0.060	-2.582	1.708
	India	22	14	1.232	0.916	-2.190	1.565	1.296	0.912	-2.472	1.744
MRPL30	Africa	12	2								
	Europe	36	9								
	India	22	7		no	outgroup	sequence				
MRPS05	Africa	18	11								
	Europe	38	19								
	India	22	31		no	outgroup	sequence				
MRPS14	Africa	22	2								
	Europe	40	1								
	India	22	1		no	outgroup	sequence				
PRF1	Africa	10	3	1.153	0.971	-1.809	1.272	0.581	0.655	-2.000	1.389
	Europe	10	1	0.739	0.867	-1.809	1.153	0.875	0.807	-2.000	1.355
	India	10	3	1.153	0.899	-2.060	1.438	1.484	0.916	-2.322	1.702

Gene	Population	n[a]	S[b]	Fu & Li's *D*	Probability[c]	Fu & Li's *D* 95% CI Lower[d]	Fu & Li's *D* 95% CI Upper[d]	Fu & Li's *F*	Probability[c]	Fu & Li's *F* 95% CI Lower[d]	Fu & Li's *F* 95% CI Upper[d]
TLR2	Africa	10	3	0.066	0.540	-1.809	1.272	-0.086	0.416	-2.322	1.509
	Europe	10	8	-0.094	0.459	-2.108	1.499	-0.429	0.369	-2.322	1.803
	India	10	9	1.593	0.965	-2.060	1.631	1.548	0.936	-2.317	1.836
TLR4	Africa	10	2	-0.410	0.346	-1.809	1.272	-0.417	0.296	-2.000	1.562
	Europe	10	2	0.989	0.772	-2.108	1.364	1.278	0.886	-2.118	1.613
	India	10	15	0.433	0.677	-1.952	1.484	0.343	0.636	-2.317	1.762
TLR5	Africa	10	0								
	Europe	10	0								
	India	10	1	0.739	0.848	-1.810	1.272	1.000	0.849	-2.000	1.484
TLR9	Africa	6	2	1.163	0.911	-1.687	1.556	1.484	0.916	-1.892	1.703
	Europe	8	1	0.793	0.787	-1.677	1.053	0.469	0.773	-1.406	1.329
	India	10	1	0.740	0.872	-1.809	1.153	0.875	0.791	-2.000	1.333
TNFAIP8L1	Africa	20	2	0.855	0.817	-2.178	1.012	0.491	0.660	-2.068	1.349
	Europe	36	10	-0.977	0.216	-2.449	1.432	-1.025	0.176	-2.400	1.678
	India	22	6	-0.279	0.387	-2.190	1.477	-0.200	0.421	-2.395	1.742
TRIF	Africa	20	24	1.312	0.954	-2.114	1.473	0.568	0.672	-2.368	1.803
	Europe	36	41	0.188	0.560	-2.425	1.498	0.177	0.592	-2.378	1.771
	India	22	22	0.923	0.842	-2.223	1.442	1.162	0.897	-2.524	1.734
TYROBP	Africa	20	8	0.799	0.761	-2.178	1.487	0.620	0.695	-2.371	1.704
	Europe	36	20	-1.193	0.136	-2.207	1.530	-1.614	0.098	-2.584	1.875
	India	22	12	0.629	0.724	-2.471	1.565	0.799	0.774	-2.391	1.731

A Gaur sequence was used as the outgroup for PRF1, TLR2, TLR4, TLR5, and TLR9 analyses and a Plains Bison was used for the remaining genes. Significant values are indicated in red. a) the number of chromosomes (i.e. twice the number of individuals) in the sample population; b) the number of segregating sites in the sample population; c) the probability of observing a lower value; d) the 95% confidence intervals as determined by coalescent simulation using DnaSP.

Table 4.7 Fu and Li's *D and *F** (tests without outgroups) values obtained for each gene in each sample population. Table continues on following page.**

Gene	Population	n[a]	S[b]	Fu & Li's *D**	Probability[c]	Fu & Li's *D** 95% CI Lower[d]	Fu & Li's *D** 95% CI Upper[d]	Fu & Li's *F**	Probability[c]	Fu & Li's *F** 95% CI Lower[d]	Fu & Li's *F** 95% CI Upper[d]
ART4	Africa	20	6	-0.154	0.394	-2.258	1.379	0.106	0.526	-2.508	1.613
	Europe	36	6	1.200	0.928	-2.373	1.363	0.922	0.768	-2.456	1.648
	India	22	33	1.510	**0.996**	**-2.466**	**1.360**	1.346	0.958	-2.586	1.537
CD2	Africa	20	13	1.098	0.906	-2.235	1.410	0.921	0.830	-2.535	1.621
	Europe	36	30	-0.370	0.313	-2.366	1.396	-0.519	0.335	-2.566	1.628
	India	22	29	0.667	0.772	-2.334	1.282	0.626	0.758	-2.475	1.493
FEZL	Africa	18	2	-0.552	0.308	-1.989	1.123	-0.522	0.252	-2.130	1.385
	Europe	36	8	-1.337	0.121	-2.375	1.262	-1.531	0.096	-2.456	1.583
	India	22	11	0.528	0.704	-2.460	1.376	0.086	0.523	-2.496	1.580
IL2	Africa	20	1	0.650	0.649	-1.540	1.006	0.520	0.719	-1.648	1.220
	Europe	36	12	-2.487	**0.016**	**-2.373**	**1.316**	-2.647	**0.017**	**-2.456**	**1.625**
	India	22	7	-1.278	0.177	-2.356	1.293	-1.311	0.124	-2.238	1.628
IL13	Africa	20	15	-2.916	**0.002**	**-2.235**	**1.342**	-3.129	**0.001**	**-2.467**	**1.541**
	Europe	36	15	-1.788	0.037	-2.025	1.441	-1.804	0.048	-2.289	1.684
	India	22	14	1.131	0.926	-2.356	1.418	1.182	0.932	-2.359	1.467
MRPL30	Africa	12	2	0.973	0.819	-1.720	1.195	0.753	0.699	-1.865	1.356
	Europe	36	9	0.142	0.363	-2.373	1.473	0.328	0.420	-2.366	1.639
	India	22	7	-1.278	0.177	-2.460	1.293	-1.338	0.139	-2.407	1.550
MRPS05	Africa	18	11	0.589	0.709	-2.367	1.404	0.933	0.83	-2.349	1.572
	Europe	38	19	1.287	0.948	-2.498	1.502	0.875	0.844	-2.253	1.670
	India	22	31	1.489	**0.997**	**-2.473**	**1.222**	1.434	**0.976**	**-2.585**	**1.422**
MRPS14	Africa	22	2	0.851	0.838	-2.109	1.240	1.195	0.906	-2.330	1.488
	Europe	40	1	0.564	0.634	-1.774	0.916	0.366	0.744	-1.835	1.075
	India	22	1	0.635	0.66	-1.575	0.992	0.479	0.757	-1.678	1.195
PRF1	Africa	10	3	-1.784	**0.024**	**-1.733**	**1.239**	-1.934	**0.015**	**-1.719**	**1.371**
	Europe	10	1	0.804	0.709	-1.587	1.154	0.898	0.753	-1.719	1.351
	India	10	3	1.154	0.886	-1.916	1.346	1.412	0.943	-1.934	1.535

Gene	Population	n[a]	S[b]	Fu & Li's *D**	Probability[c]	Fu & Li's *D** 95% CI Lower[d]	Fu & Li's *D** 95% CI Upper[d]	Fu & Li's *F**	Probability[c]	Fu & Li's *F** 95% CI Lower[d]	Fu & Li's *F** 95% CI Upper[d]
TLR2	Africa	10	3	0.174	0.640	-1.784	1.239	0.010	0.545	-1.934	1.467
	Europe	10	8	-1.284	0.129	-2.010	1.346	-1.389	0.121	-2.179	1.660
	India	10	9	1.435	**0.985**	**-2.025**	**1.411**	1.341	0.921	-2.179	1.615
TLR4	Africa	10	2	-0.280	0.375	-1.784	1.239	-0.287	0.345	-1.934	1.412
	Europe	10	2	1.026	0.904	-1.916	1.300	1.255	0.895	-1.934	1.533
	India	10	15	-0.028	0.469	-1.935	1.373	-0.054	0.467	-2.250	1.553
TLR5	Africa	10	0								
	Europe	10	0								
	India	10	1	0.804	0.682	-1.587	1.154	1.026	0.825	-1.719	1.337
TLR9	Africa	6	2	1.280	0.871	-1.325	1.467	1.434	0.926	-1.396	1.582
	Europe	8	1	-1.126	0.167	-1.410	1.111	-1.204	0.183	-1.514	1.297
	India	10	1	0.804	0.695	-1.587	1.154	0.898	0.754	-1.720	1.351
TNFAIP8L1	Africa	20	2	0.866	0.810	-2.053	1.108	0.488	0.635	-2.188	1.327
	Europe	36	10	-0.860	0.203	-2.507	1.404	-0.919	0.194	-2.323	1.664
	India	22	6	-0.917	0.168	-2.460	1.338	-0.727	0.273	-2.407	1.581
TRIF	Africa	20	24	1.164	0.923	-2.474	1.410	0.513	0.681	-2.452	1.563
	Europe	36	41	0.212	0.586	-2.520	1.262	0.195	0.566	-2.634	1.616
	India	22	22	0.848	0.834	-2.560	1.321	1.052	0.890	-2.426	1.482
TYROBP	Africa	20	8	0.777	0.736	-2.386	1.379	0.598	0.707	-2.535	1.554
	Europe	36	20	-1.344	0.094	-2.206	1.441	-1.688	0.077	-2.566	1.642
	India	22	12	0.614	0.738	-2.236	1.394	0.759	0.787	-2.440	1.659

Significant values are indicated in red. a) the number of chromosomes (i.e. twice the number of individuals) in the sample population; b) the number of segregating sites in the sample population; c) the probability of observing a lower value; d) the 95% confidence intervals as determined by coalescent simulation using DnaSP.

4.3.5.3 Fay and Wu's H

Fay and Wu's H (Fay and Wu 2000) is a statistic that compares two different estimators of θ: π and θ_H. θ_H is weighted towards high frequency derived alleles, thus a negative value for Fay and Wu's H indicated an excess of high frequency derived variants and a positive value indicates a dearth of high frequency derived variants.

In order to calculate Fay and Wu's H an outgroup sequence must be available to determine the state of the variants (ancestral or derived). Fay and Wu's H was calculated using DnaSP for all genes for which an outgroup sequence was available. For PRF1, TLR2, TLR4, TLR5 and TLR9, the outgroup sequence was Gaur (*Bos gaurus*). For ART4, CD2, FEZL, IL2, IL13, TNFAIP8L1, TRIF and TYROBP, the outgroup sequence was Plains Bison (*Bison bison*). The results are outlined in Table 4.8

4.3.5.4 Fu's Fs

Fu's F_S (Fu 1997) is based on the probability of observing a greater number of alleles in the population then expected under neutrality. Fu's F_S is a one sided test, where a significant negative result indicates an excess of recent mutations. An excess of recent mutations could be the result of positive selection, or could be indicative of a recent population expansion. The results of Fu's F_S test, as calculated in DnaSP, are presented in Table 4.9.

Fu's F_S test showed only four significant results, one of which was marginal (TYROBP in European *Bos taurus*, p=0.04). IL13 was significantly negative in both the European *Bos taurus* (p=0.018) and the African *Bos taurus* (p=0.005). IL2 was significantly negative in the European *Bos taurus* only (p=0.014). These significantly negative results indicate an excess of high frequency derived alleles in these genes in these populations.

Table 4.8 Fay and Wu's *H* values obtained for each gene in each sample population. Table continues on following page.

Gene	Population	n[a]	S[b]	Fay & Wu's *H*	Probability[c]	Fay & Wu's *H* 95% CI Lower[d]	Fay & Wu's *H* 95% CI Upper[d]
ART4	Africa	20	6	0.779	0.674	-4.253	1.811
	Europe	36	6	0.397	0.578	-3.683	1.460
	India	22	33	-2.338	0.223	-17.697	6.649
CD2	Africa	20	13	-1.200	0.199	-8.653	2.947
	Europe	36	30	-9.511	0.042	-12.219	4.337
	India	22	29	-7.775	0.094	-16.277	6.355
FEZL	Africa	18	2	0.366	0.797	-1.778	0.784
	Europe	36	8	0.898	0.931	-3.133	1.194
	India	22	11	1.784	0.961	-3.861	2.043
IL2	Africa	20	1	-1.253	0.029	-1.421	0.505
	Europe	36	12	-5.680	**0.013**	-3.810	1.292
	India	22	7	0.156	0.342	-3.706	1.558
IL13	Africa	20	15	1.621	0.975	-4.505	1.632
	Europe	36	15	1.756	0.925	-5.317	2.162
	India	22	14	2.130	0.846	-8.225	3.411
MRPL30	Africa	12	2				
	Europe	36	9				
	India	22	7	No	outgroup	sequence	
MRPS05	Africa	18	11				
	Europe	38	19				
	India	22	31	No	outgroup	sequence	
MRPS14	Africa	22	2				
	Europe	40	1				
	India	22	1	No	outgroup	sequence	
PRF1	Africa	10	3	-4.800	**0.003**	**-1.600**	**0.889**
	Europe	10	1	0.267	0.795	-1.689	0.711
	India	10	3	0.622	0.716	-3.467	1.422

Gene	Population	n[a]	S[b]	Fay & Wu's *H*	Probability[c]	Fay & Wu's *H* 95% CI Lower[d]	Fay & Wu's *H* 95% CI Upper[d]
TLR2	Africa	10	3	0.711	0.893	-2.667	1.067
	Europe	10	8	-4.000	0.038	-4.889	1.867
	India	10	9	0.800	0.603	-6.756	2.578
TLR4	Africa	10	2	0.444	0.782	-1.867	0.889
	Europe	10	2	-0.622	0.157	-3.022	1.156
	India	10	15	-6.311	0.042	-8.089	3.822
TLR5	Africa	10	0				
	Europe	10	0				
	India	10	1	0.178	0.688	-1.600	0.800
TLR9	Africa	6	2	0.000	0.465	-2.400	1.333
	Europe	8	1	-1.500	**0.023**	**-1.286**	**0.571**
	India	10	1	0.267	0.796	-1.600	0.711
TNFAIP8L1	Africa	20	2	0.337	0.837	-1.705	0.611
	Europe	36	10	-0.838	0.170	-4.492	1.724
	India	22	6	-0.987	0.158	-4.485	1.541
TRIF	Africa	20	24	-19.179	**0.001**	**-9.126**	**3.263**
	Europe	36	41	-7.137	0.103	-18.435	7.178
	India	22	22	-5.576	0.110	-15.290	5.385
TYROBP	Africa	20	8	-3.779	0.044	-4.916	1.968
	Europe	36	20	-7.356	**0.010**	**-4.959**	**2.232**
	India	22	12	1.887	0.847	-8.710	3.152

a) the number of chromosomes (i.e. twice the number of individuals) in the sample population; b) the number of segregating sites in the sample population; c) the probability of observing a lower value; d) the 95% confidence intervals as determined by coalescent simulation using DnaSP

Table 4.9 Fu's F_S, observed value and probability for each gene in each sample population. Table continues on following page.

Gene	Population	n[a]	Number of haplotypes	Fu's F_S	Probability[b]	Fu's F_S 95% CI Lower[c]
ART4	Africa	20	6	-0.168	0.468	-3.102
	Europe	36	6	-0.481	0.422	-3.490
	India	22	16	-3.026	0.122	-5.155
CD2	Africa	20	5	2.696	0.899	-4.148
	Europe	36	10	1.765	0.790	-6.225
	India	22	10	1.297	0.743	-5.048
FEZL	Africa	18	3	-0.161	0.415	-1.744
	Europe	36	7	-2.090	0.098	-3.188
	India	22	9	-2.751	0.056	-3.680
IL2	Africa	20	2	0.381	0.520	-1.863
	Europe	36	9	-3.667	**0.014**	-3.188
	India	22	7	-2.008	0.091	-2.982
IL13	Africa	20	9	-3.877	**0.005**	-3.088
	Europe	36	13	-4.530	**0.018**	-4.338
	India	22	12	-2.505	0.123	-4.381
MRPL30	Africa	12	2	1.384	0.887	-1.611
	Europe	36	7	0.706	0.656	-4.246
	India	22	5	-0.105	0.487	-2.889
MRPS05	Africa	18	10	-1.721	0.221	-4.198
	Europe	38	10	0.232	0.573	-5.158
	India	22	16	-2.997	0.114	-5.497
MRPS14	Africa	22	4	-0.096	0.501	-2.367
	Europe	40	2	0.129	0.469	-1.548
	India	22	2	0.303	0.613	-1.310
PRF1	Africa	10	2	1.225	0.845	-1.430
	Europe	10	2	0.818	0.696	-1.164
	India	10	3	1.358	0.811	-2.291

Gene	Population	n[a]	Number of haplotypes	Fu's F_S	Probability[b]	Fu's F_S 95% CI Lower[c]
TLR2	Africa	10	4	-1.071	0.130	-1.833
	Europe	10	5	-0.479	0.332	-2.442
	India	10	4	1.958	0.842	-2.845
TLR4	Africa	10	3	-0.272	0.373	-1.430
	Europe	10	3	0.520	0.643	-1.944
	India	10	5	1.660	0.792	-3.349
TLR5	Africa	10	1			
	Europe	10	1			
	India	10	2	1.029	0.766	-1.176
TLR9	Africa	6	2	1.938	0.927	-1.909
	Europe	8	2	-0.182	0.453	-0.844
	India	10	2	0.818	0.694	-1.164
TNFAIP8L1	Africa	20	3	-0.724	0.305	-1.863
	Europe	36	5	1.468	0.802	-3.695
	India	22	4	1.450	0.829	-2.947
TRIF	Africa	20	5	3.506	0.940	-4.516
	Europe	36	11	3.694	0.914	-6.340
	India	22	7	3.869	0.936	-5.074
TYROBP	Africa	20	6	-0.050	0.525	-3.227
	Europe	36	12	-3.455	0.040	-4.212
	India	22	11	-2.192	0.131	-4.299

a) the number of chromosomes (i.e. twice the number of individuals) in the sample population; b) the probability of obtaining a lower value than the observed value; c) the 95% lower confidence interval as determined by coalescent simulation using DnaSP.

4.3.6 Sliding window Analyses

In addition to calculating each neutrality test statistic over the whole region sequenced, a sliding window analysis was carried out for i) nucleotide diversity (π) per site, ii) Tajima's *D*, iii) Fu and Li's *D** and *F** and iv) divergence between continental populations. For all sliding window analyses a window size of 150 bp and a slide of 25 bp was used. Given that selection acts on only one or a few sites, sliding window analyses are a useful tool to explore the way the test statistics vary across the region and may highlight signals that are not apparent when the whole gene is considered. Additionally, recombinant haplotypes may be a confounding factor in interpreting the results of summary statistics calculated over the whole gene. Such an effect is less likely to influence more limited windows.

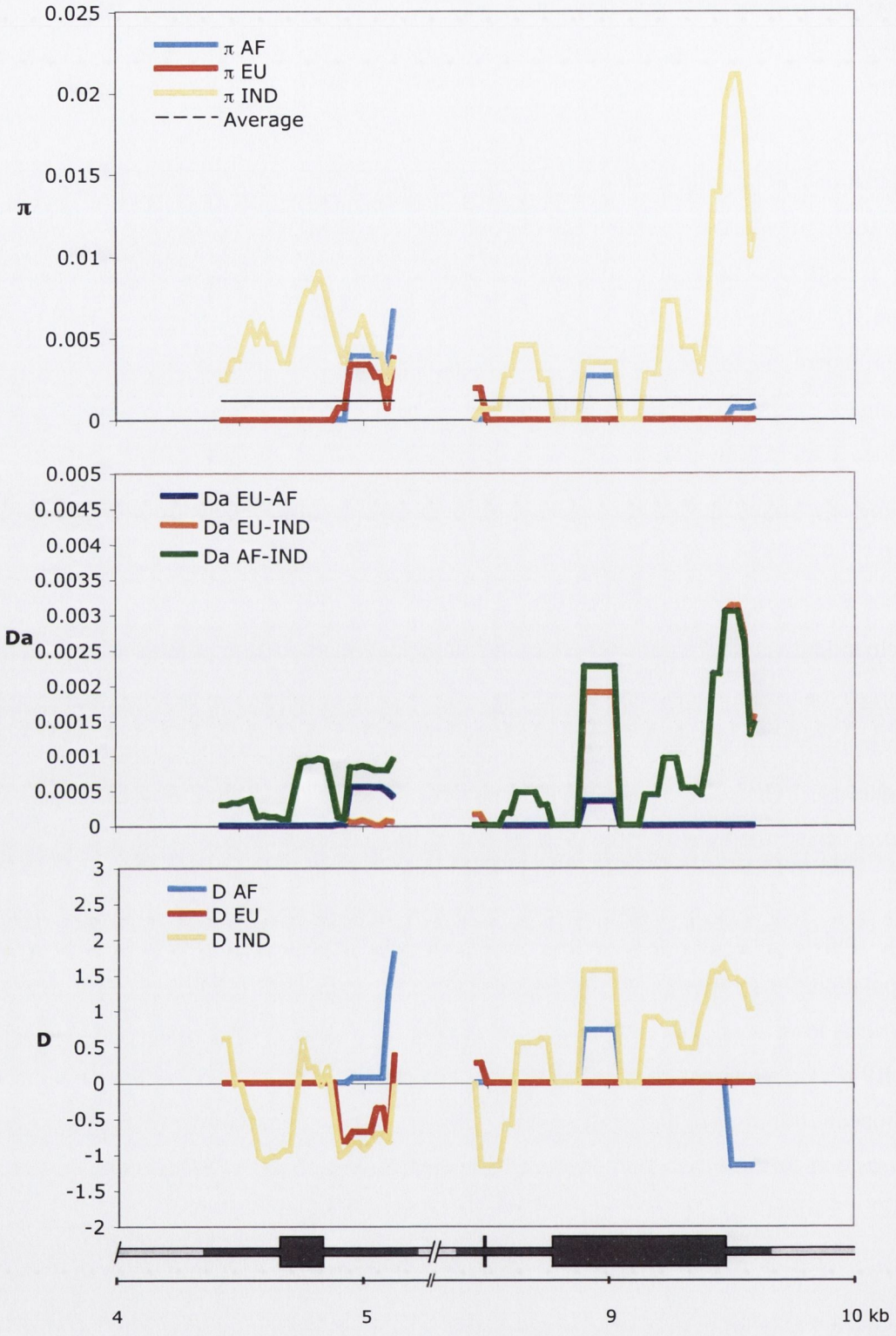
Figure 4.7 shows the results of the sliding window analysis for each gene excluding TLR5, which had only one polymorphic site in *Bos indicus* and no polymorphic sites in either *Bos taurus* population, and TLR7, which had no polymorphic sites. The average nucleotide diversity (π per site) for each continental population is plotted along with the average π across all populations and genes ($\pi = 0.00117$). There is a trend towards higher values of π in *Bos indicus*, consistent with previous results (haplotype diversity, SNP frequency). There are some notable exceptions. TRIF shows higher π in European *Bos taurus* across the length of the gene as does MRPL30, particularly at the 3' end of the gene.

The middle plot of each figure shows the divergence (D_a) between continental populations. As expected, the divergence between European *Bos taurus* and African *Bos taurus* is lower than the divergence between either *Bos taurus* population and *Bos indicus*. TLR9 is exception, with higher levels of divergence between African *Bos taurus* and both European *Bos taurus* and *Bos indicus*, and lower divergence between European *Bos taurus* and *Bos indicus*. Low levels of divergence were observed for all comparisons in FEZL. Tajima's *D* is plotted in the lower graph of each figure.

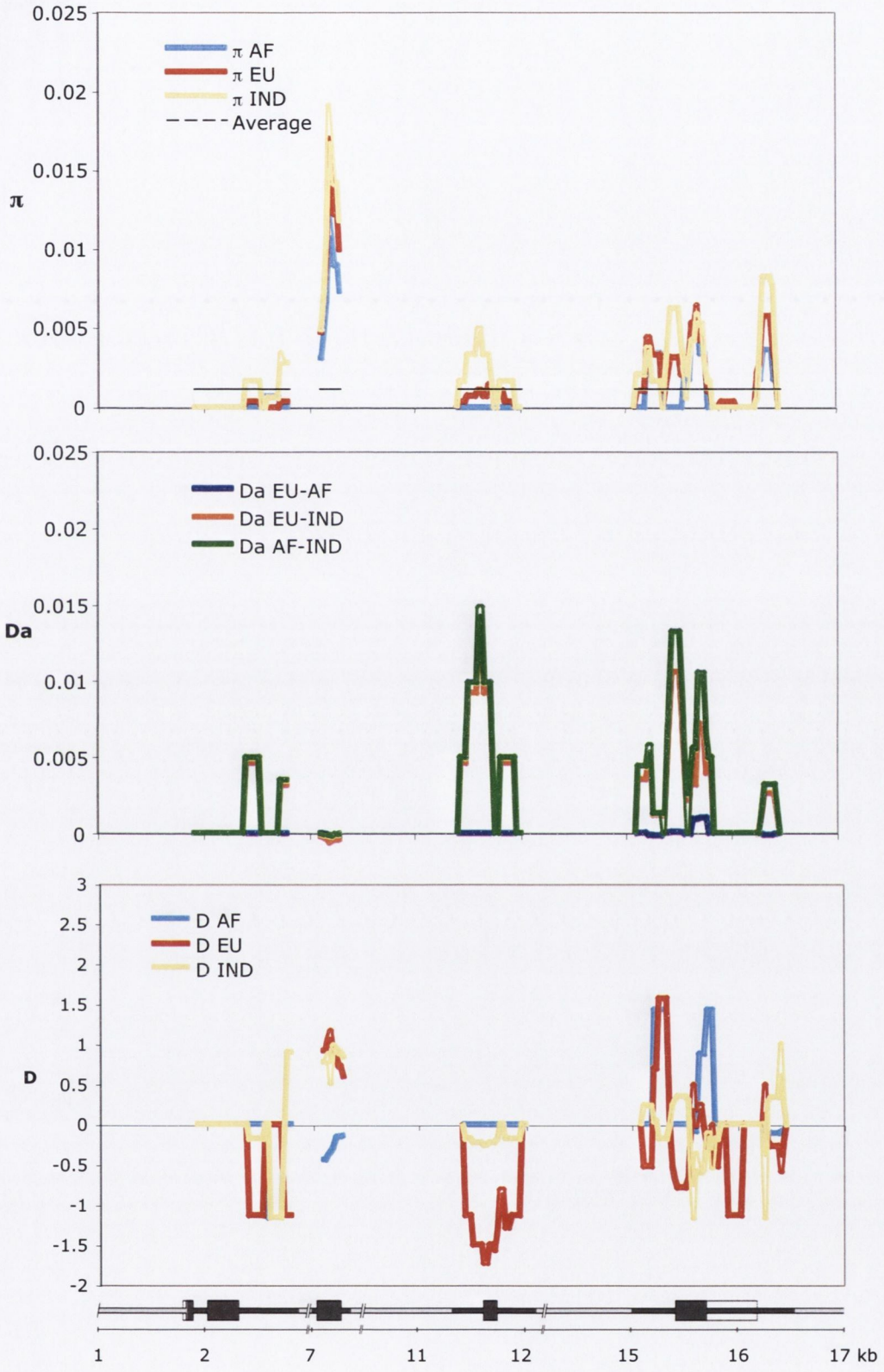
Figure 4.7 Variation in π, divergence between continental populations (D_a) and Tajima's D across each gene.

(On following pages.)

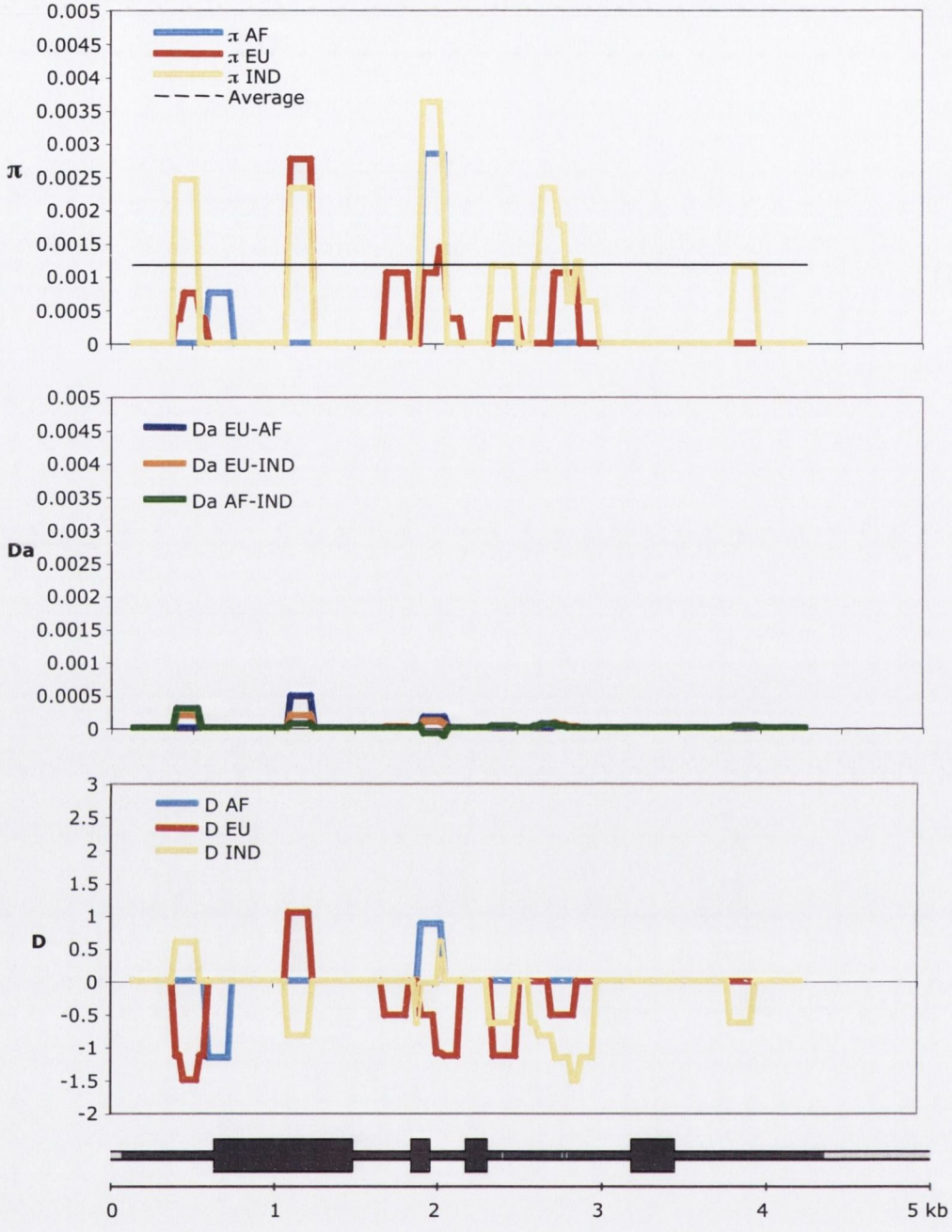
a) ART4


0.025
0.02
0.015
π
0.01
0.005
0
π AF
π EU
π IND
Average
0.005
0.0045
0.004
0.0035
0.003
Da
0.0025
0.002
0.0015
0.001
0.0005
0
Da EU-AF
Da EU-IND
Da AF-IND
3
2.5
2
1.5
1
D
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
4
5
9
10 kb

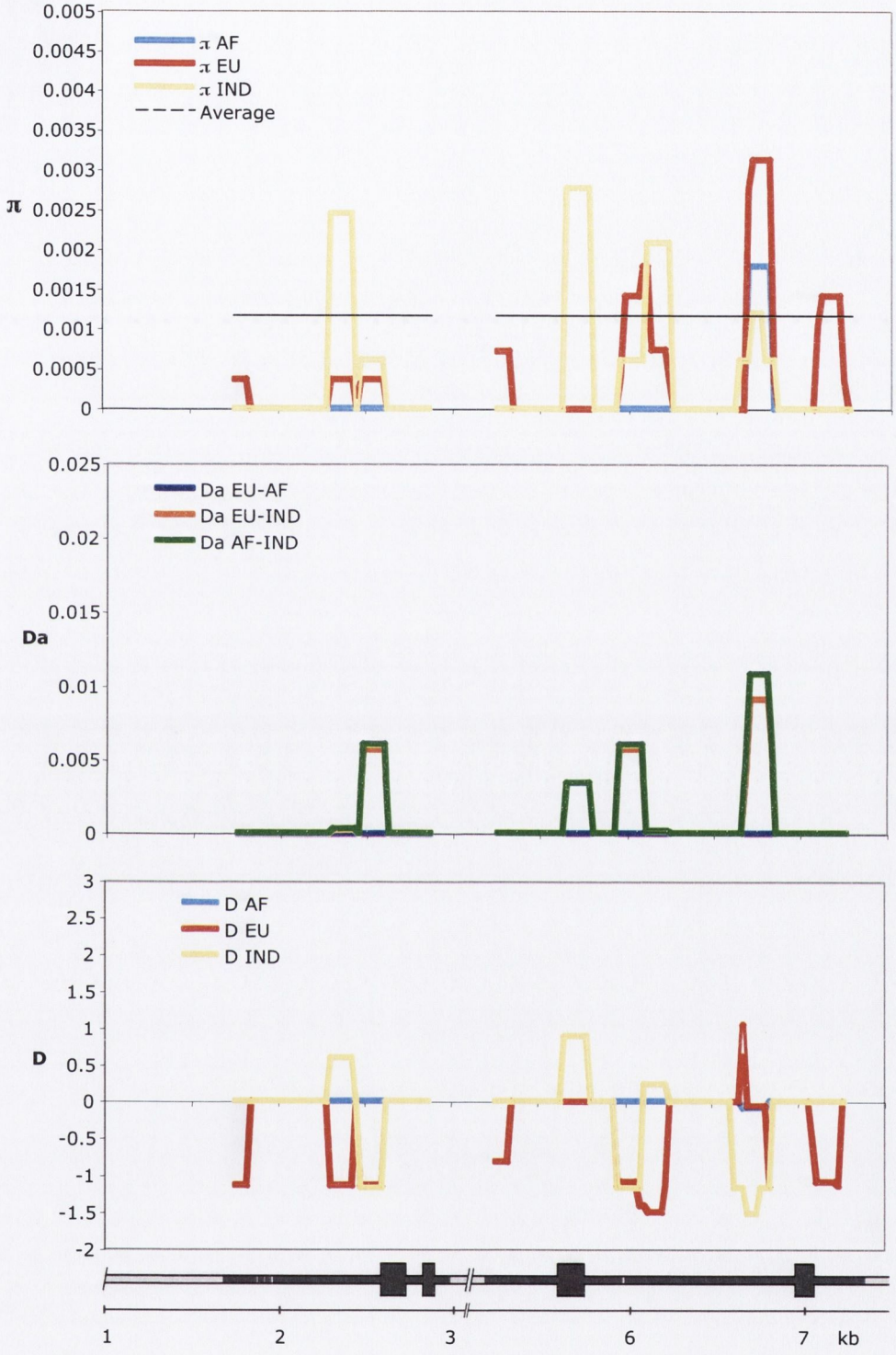
b) CD2


π AF
π EU
π IND
Average
π
0.025
0.02
0.015
0.01
0.005
0
Da EU-AF
Da EU-IND
Da AF-IND
Da
D AF
D EU
D IND
D
3
2.5
2
1.5
1
0.5
-0.5
-1
-1.5
-2
1
2
7
11
12
15
16
17 kb

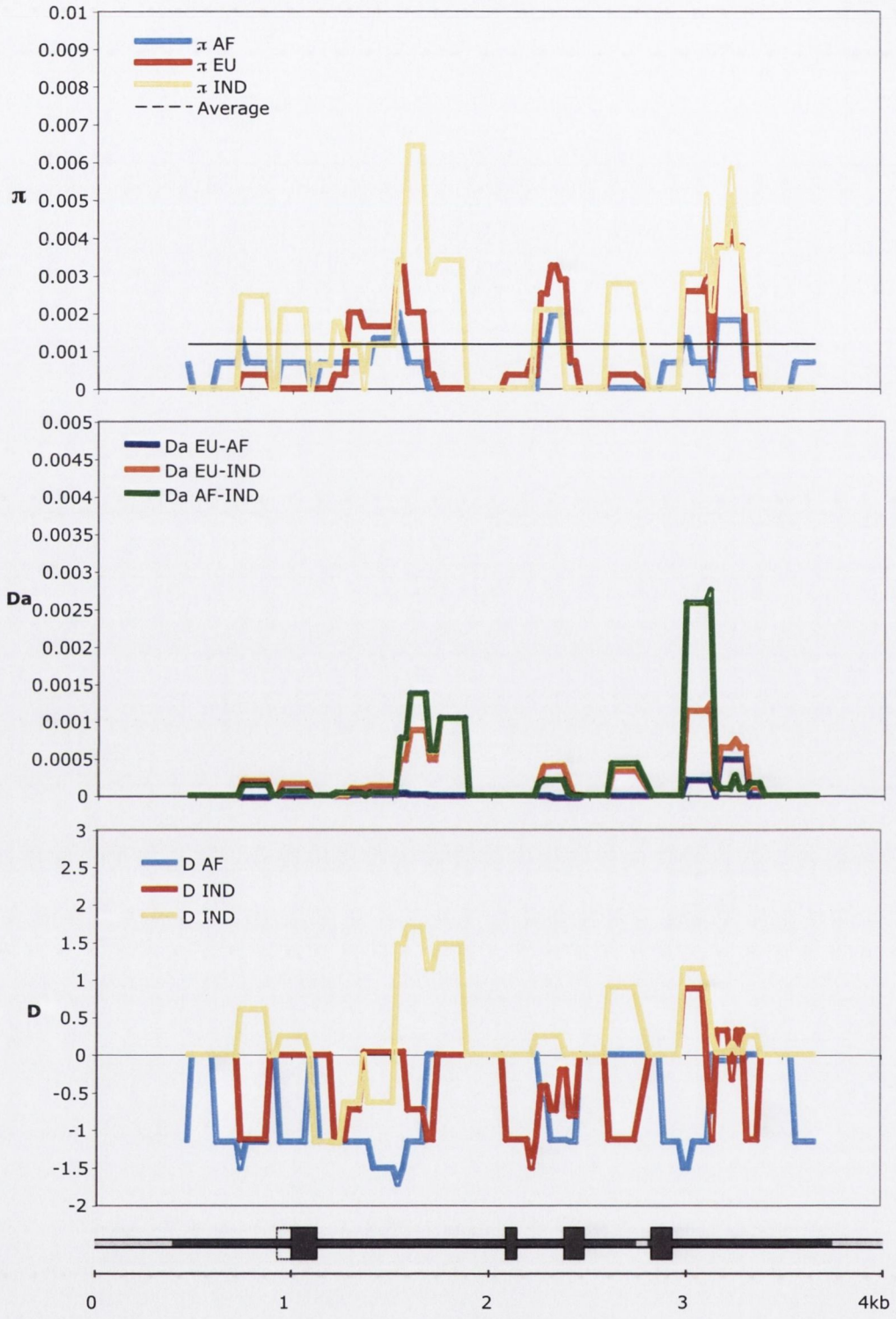
c) FEZL


π
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π AF
π EU
π IND
Average
Da
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
0
1
2
3
4
5 kb

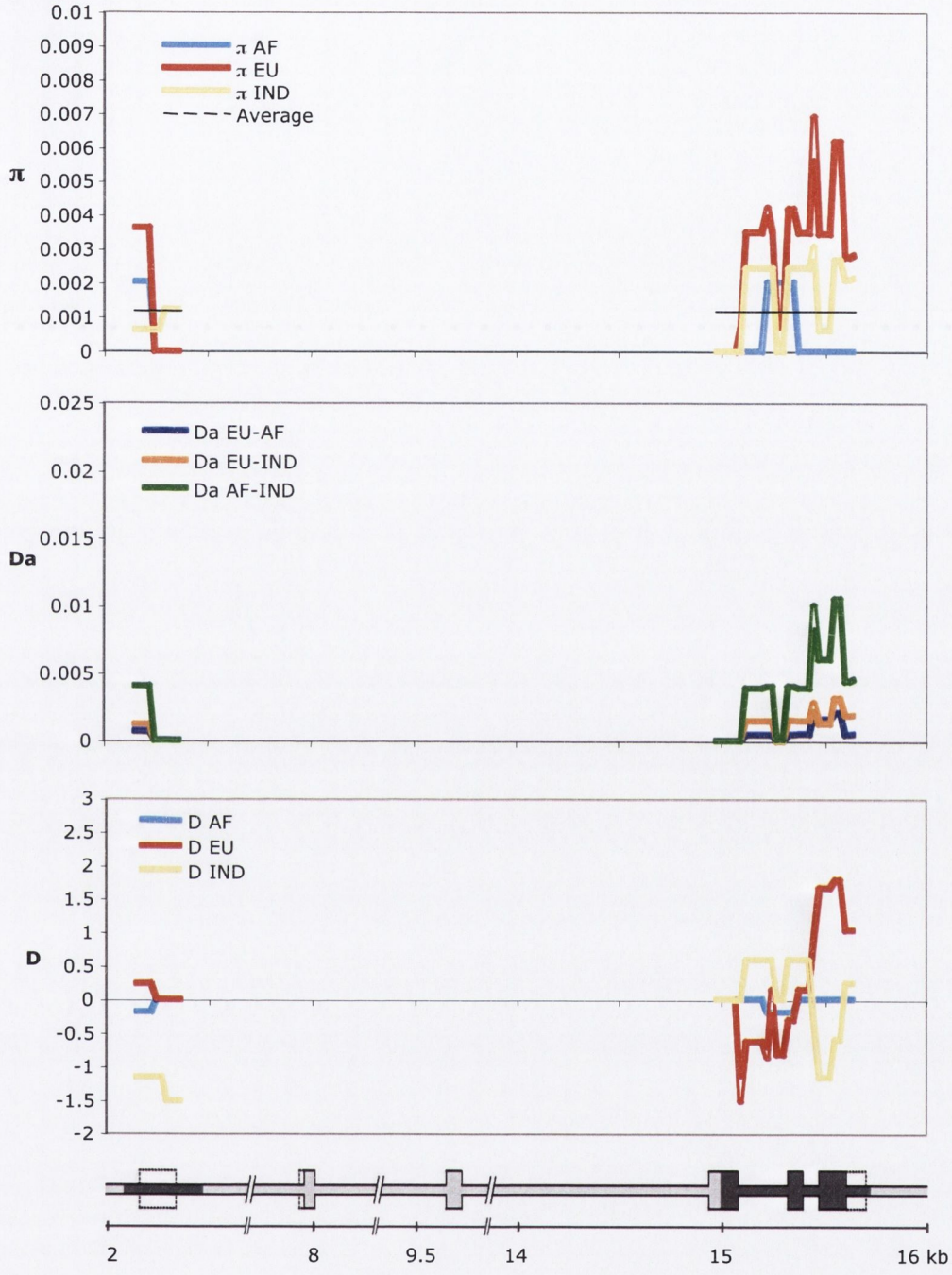
d)IL2


π AF
π EU
π IND
Average
π
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
Da EU-AF
Da EU-IND
Da AF-IND
Da
0.025
0.02
0.015
0.01
0.005
0
D AF
D EU
D IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
1
2
3
6
7
kb

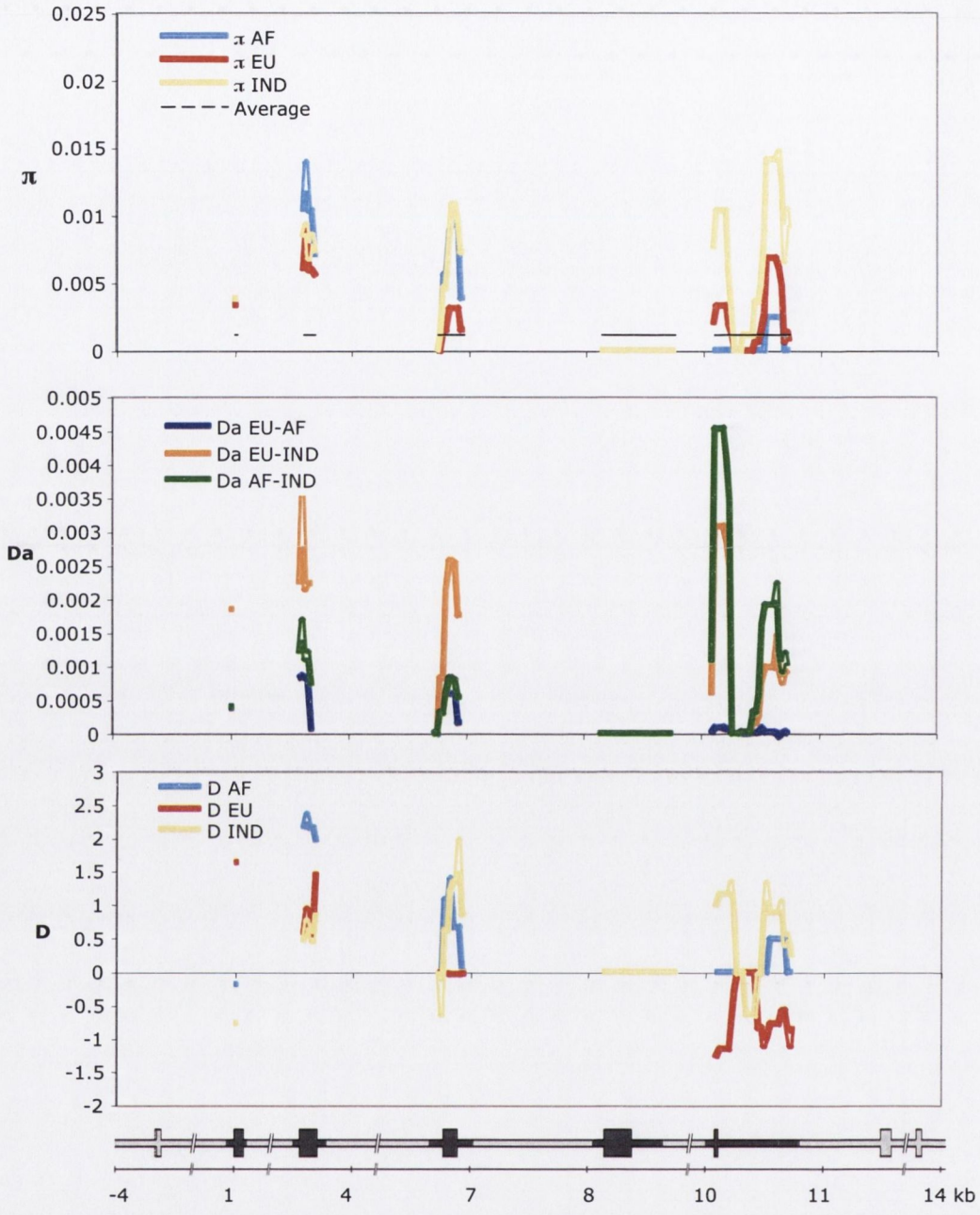
e) IL13


π AF
π EU
π IND
Average
π
Da EU-AF
Da EU-IND
Da AF-IND
Da
D AF
D IND
D IND
D
0
1
2
3
4kb

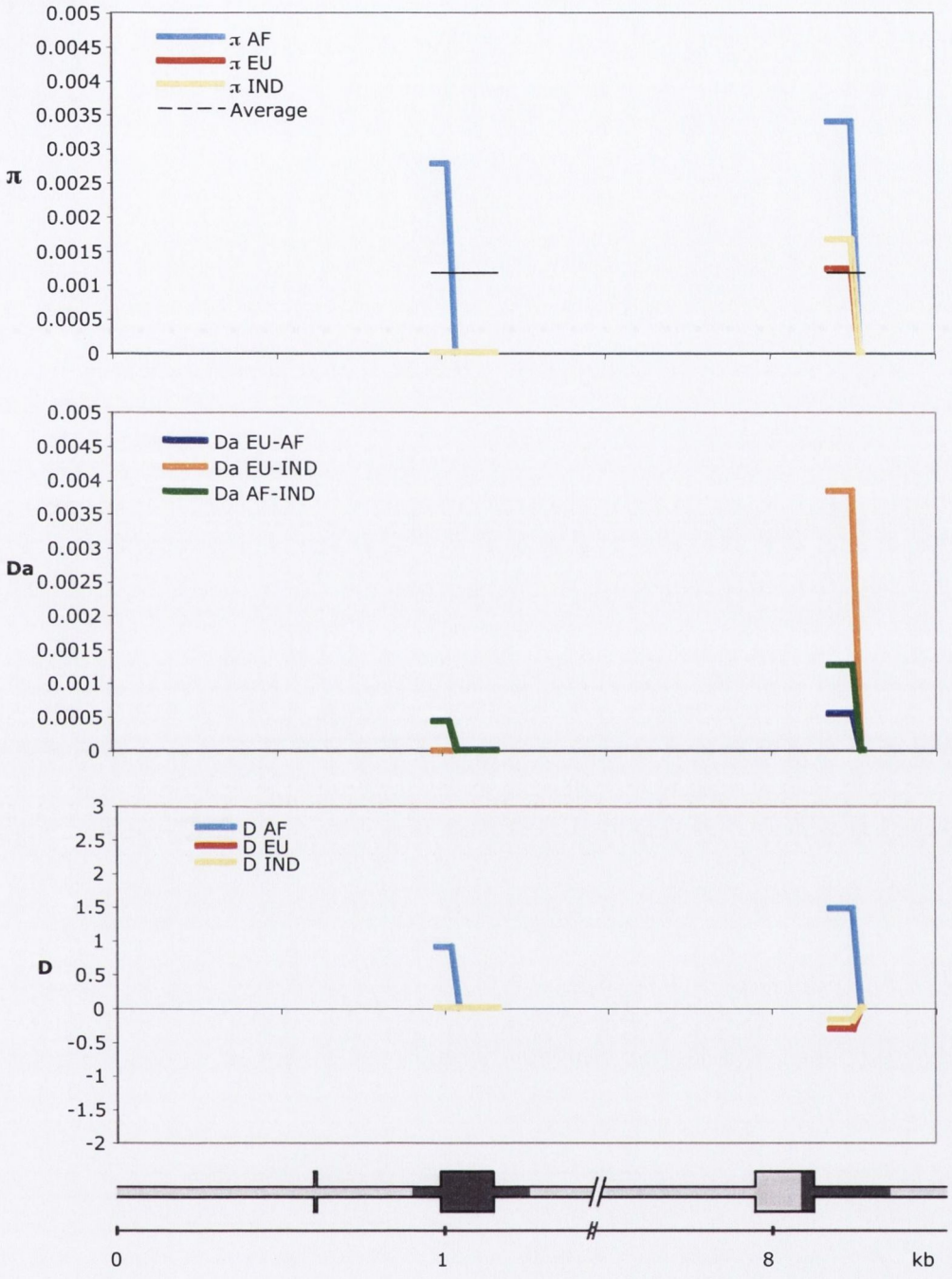
f) MRPL30


π
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
π AF
π EU
π IND
Average
Da
0.025
0.02
0.015
0.01
0.005
0
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
2
8
9.5
14
15
16 kb

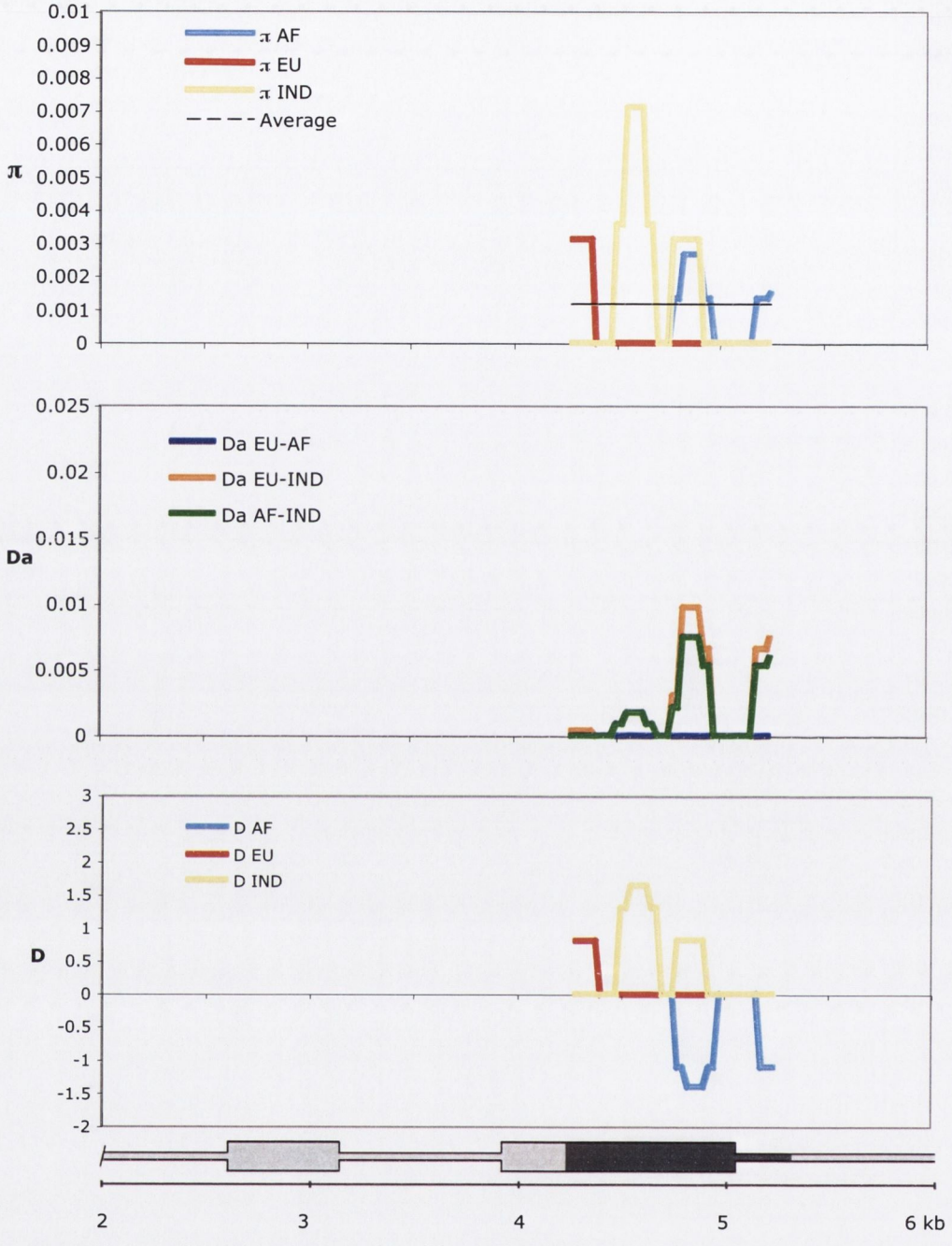
g) MRPS05


π AF
π EU
π IND
Average
π
Da EU-AF
Da EU-IND
Da AF-IND
Da
D AF
D EU
D IND
D
-4
1
4
7
8
10
11
14 kb

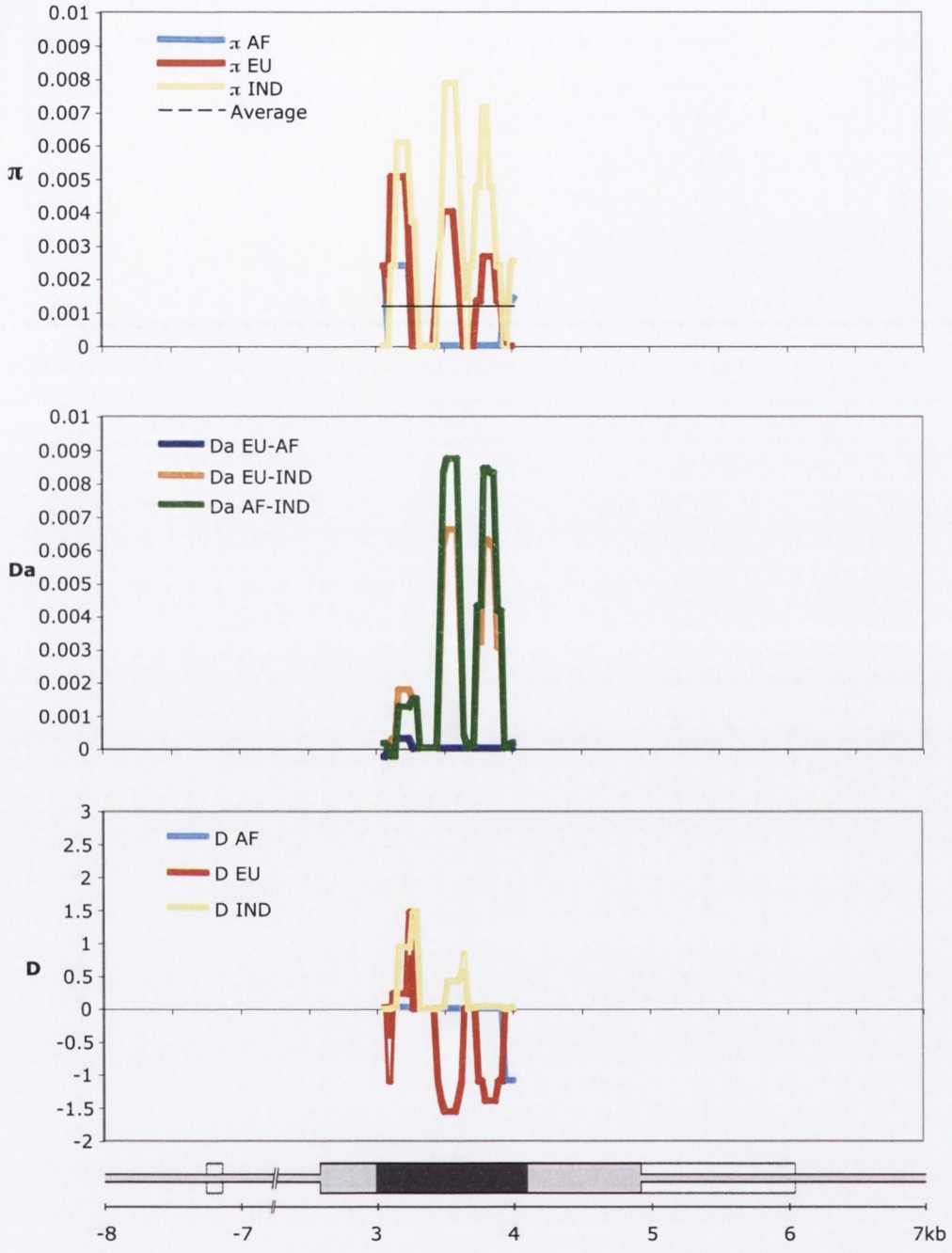
h) MRPS14


π
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
π AF
π EU
π IND
Average
Da
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
0
1
8
kb

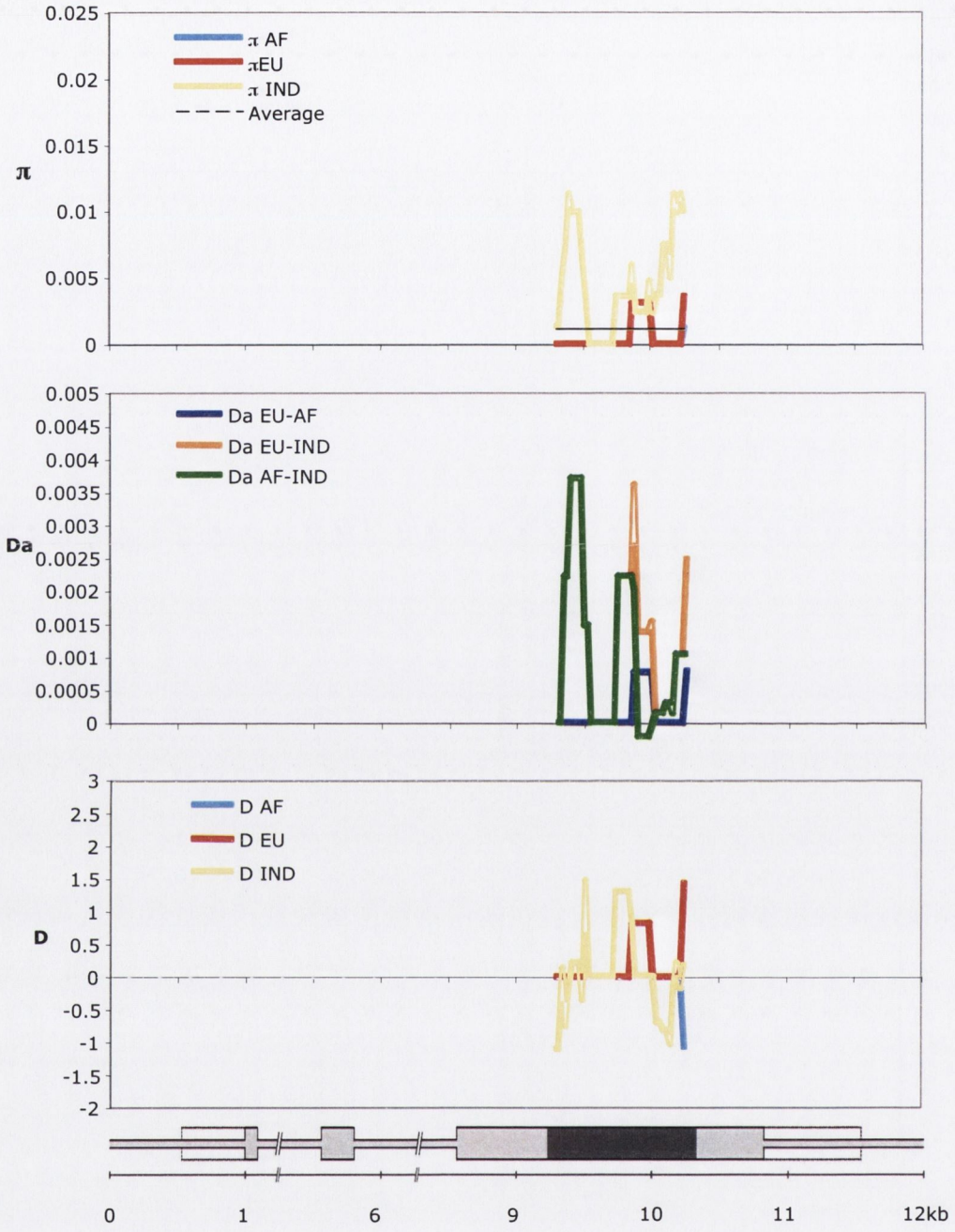
i) PRF1


π
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
π AF
π EU
π IND
Average
Da
0.025
0.02
0.015
0.01
0.005
0
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
2
3
4
5
6 kb

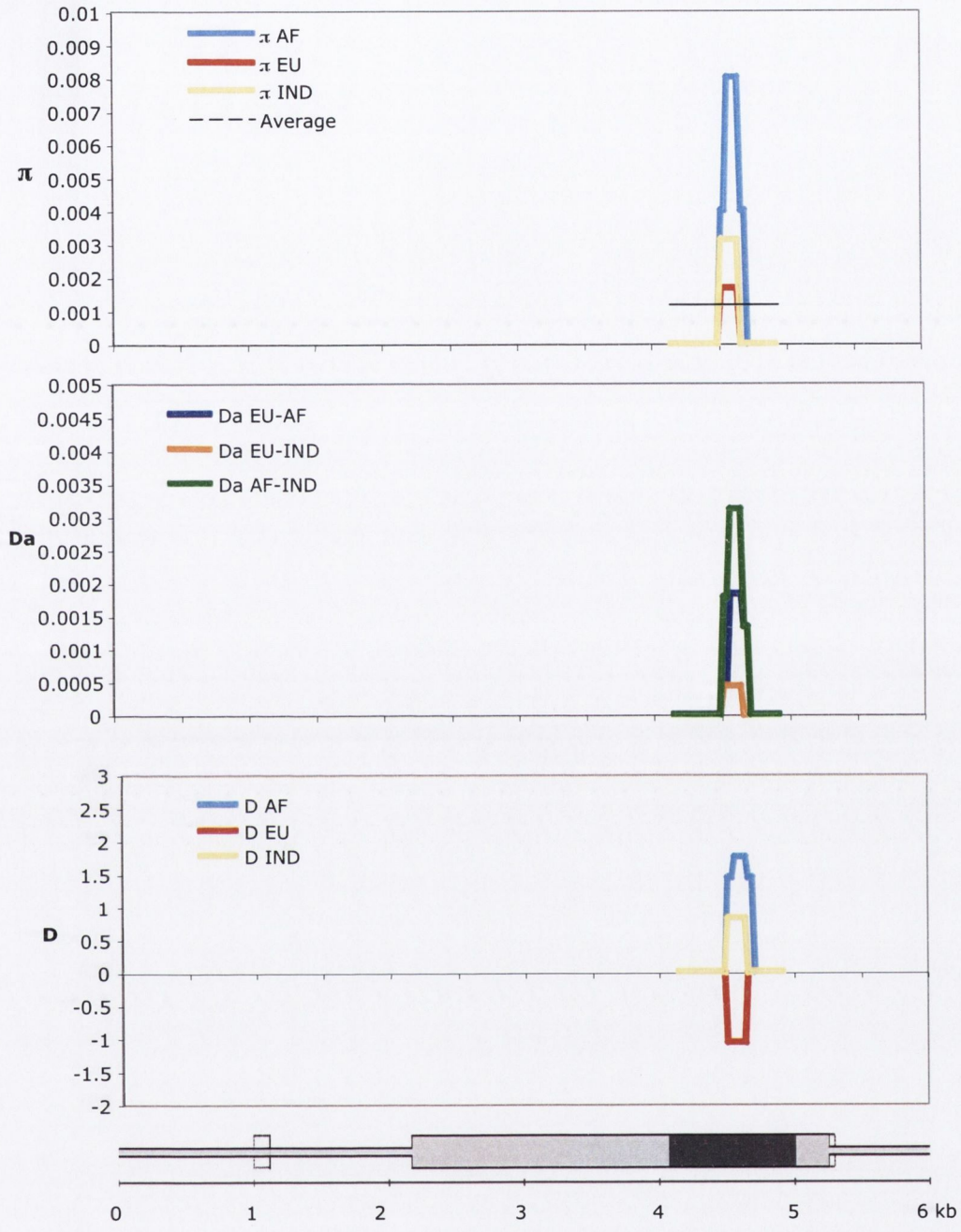
j) TLR2


π
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
π AF
π EU
π IND
Average
Da
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
-8
-7
3
4
5
6
7kb

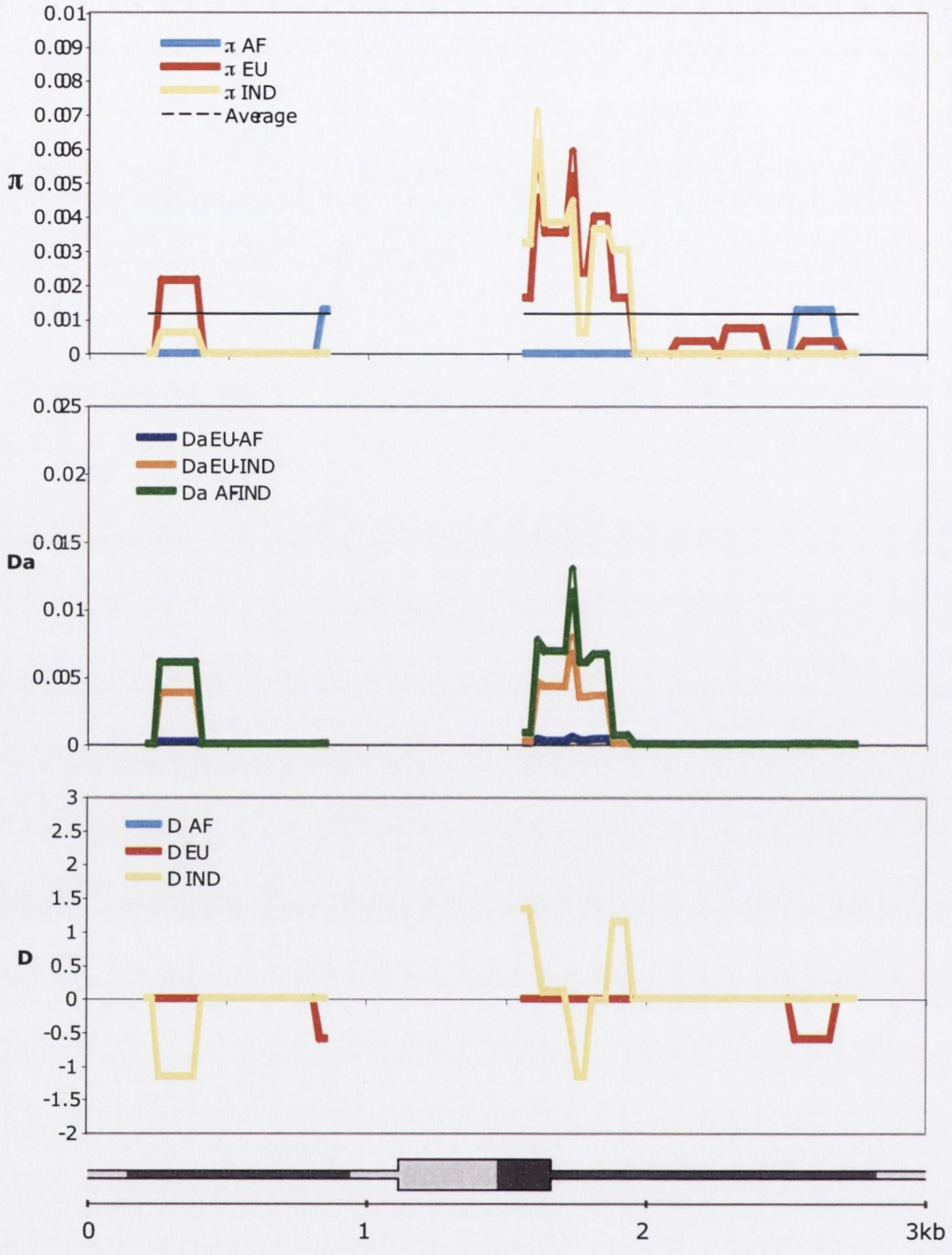
k) TLR4


0.025
0.02
0.015
π
0.01
0.005
0
π AF
πEU
π IND
Average
0.005
0.0045
0.004
0.0035
0.003
Da
0.0025
0.002
0.0015
0.001
0.0005
0
Da EU-AF
Da EU-IND
Da AF-IND
3
2.5
2
1.5
1
D
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
0
1
6
9
10
11
12kb

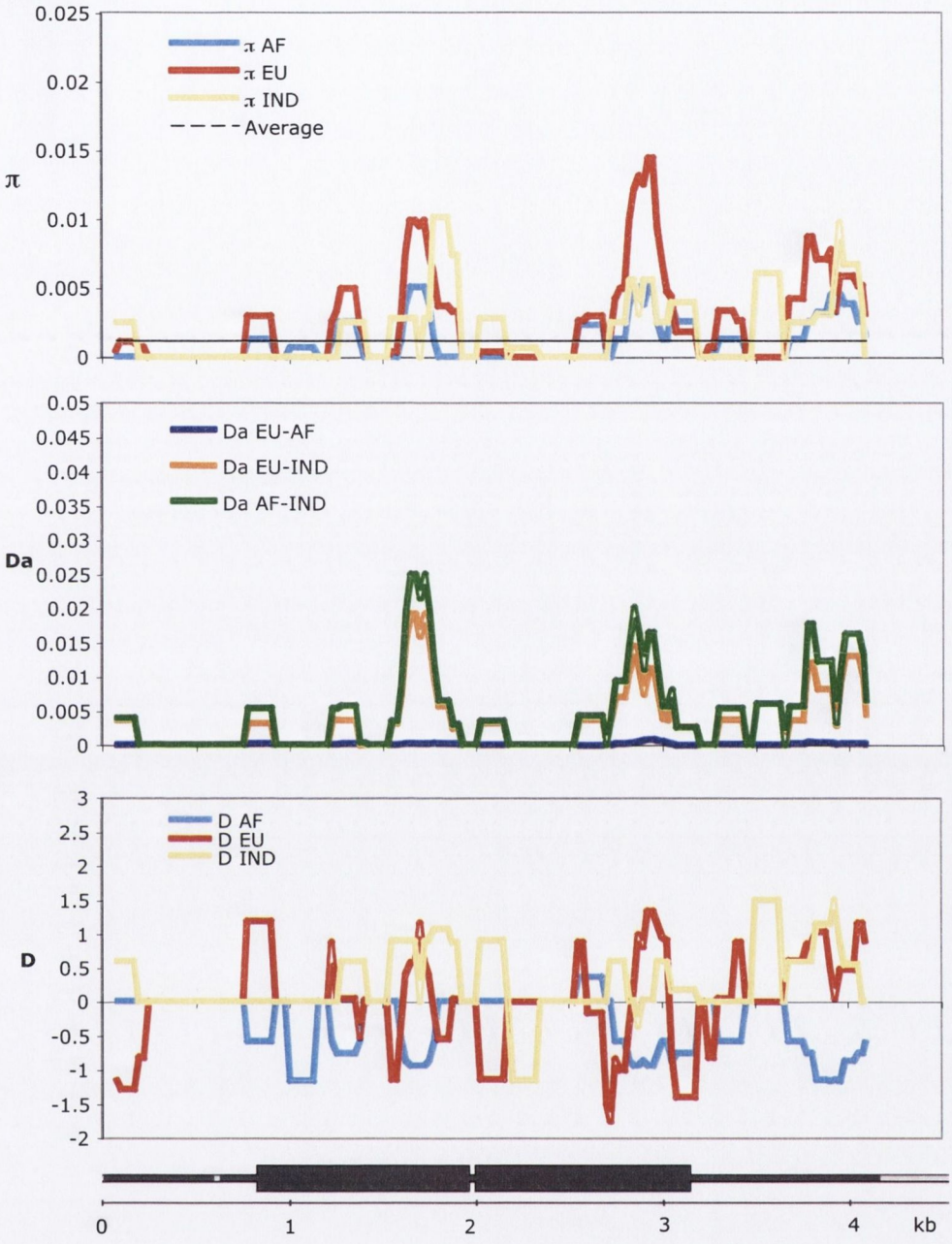
l) TLR9


π
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
π AF
π EU
π IND
Average
Da
0.005
0.0045
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
0
1
2
3
4
5
6 kb

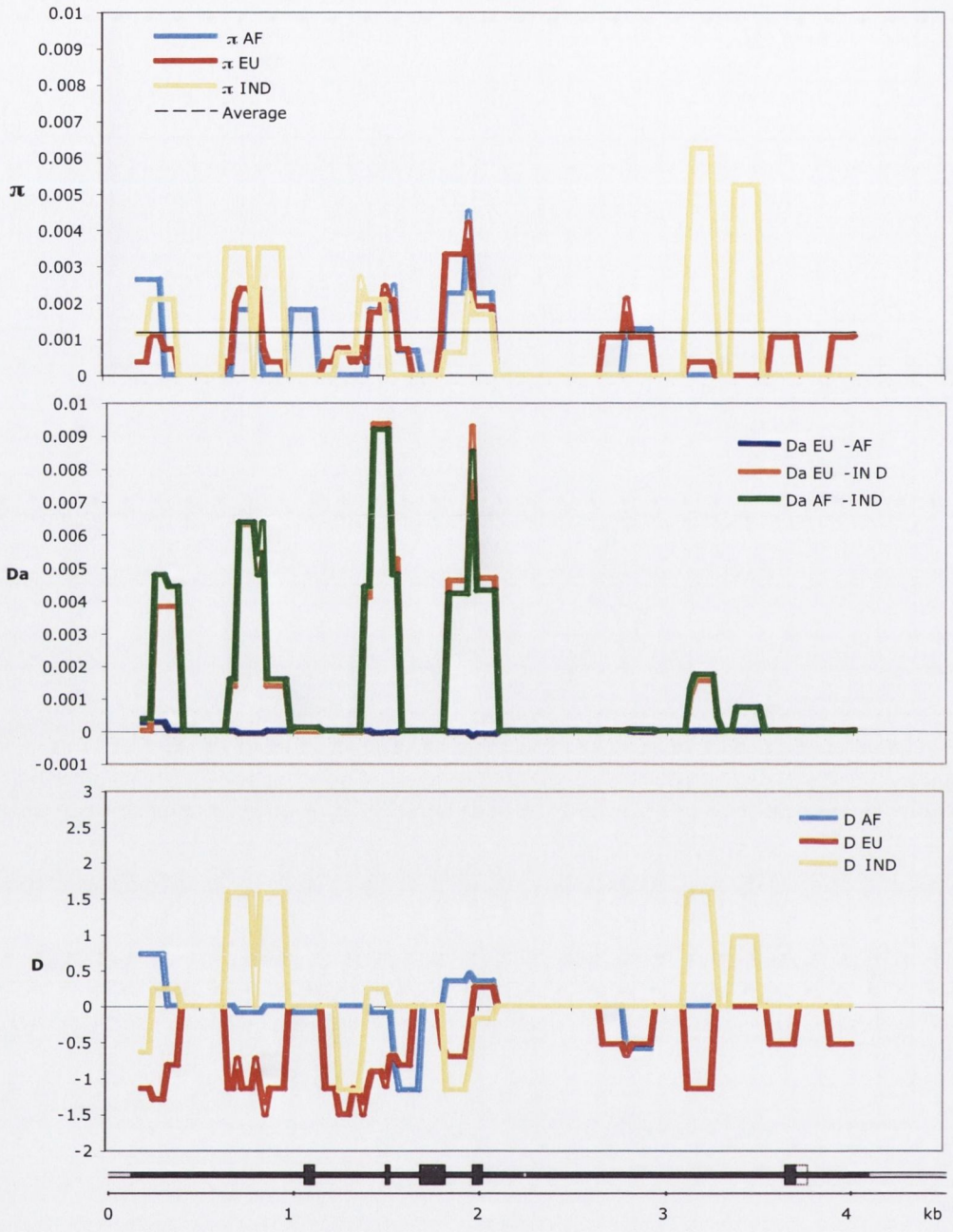
m) TNFAIP8L1


π
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
π AF
π EU
π IND
Average
Da
0.025
0.02
0.015
0.01
0.005
0
DaEU-AF
DaEU-IND
Da AFIND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
0
1
2
3kb

n) TRIF


π
0.025
0.02
0.015
0.01
0.005
0
π AF
π EU
π IND
Average
Da
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Da EU-AF
Da EU-IND
Da AF-IND
D
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
D AF
D EU
D IND
0
1
2
3
4
kb

o) TYROBP


π AF
π EU
π IND
Average
π
Da EU -AF
Da EU -IN D
Da AF -IND
Da
D AF
D EU
D IND
D
kb

4.3.7 Assessing significance of neutrality tests with coalescent simulations

4.3.7.1 MS analyses (whole gene analyses)

Patterns of polymorphism in populations are influenced not only by selection but also by the demographic history of the population. In order to test the significance of the values obtained from the observed data (i.e. if demographic history can be determined to be the cause of the observed value rather than deviations from neutrality), confidence intervals were determined by simulating data under a number of different demographic histories. Tajima's *D* and Fay and Wu's *H* were calculated in simulated populations generated using MS (Hudson 2002) to have the same properties (number of segregating sites, number of samples and sequence length) as the study dataset. These simulated datasets were generated under two different neutral models: standard neutral model (SNM), in which the population size was constant; and an expansion model (EXP1) in which the ancestral population size decreased suddenly 10,000 years ago and immediately began to increase in size, and continued to do so until the present day (Figure 4.2). The results of these analyses are presented in Table 4.10.

The observed values of Tajima's *D* and Fay and Wu's *H* were compared to the 95% confidence intervals calculated from the simulated data. Significant results, results where the observed value falls outside the 95% confidence interval, are highlighted in bold red. Values in red but not bold indicate that the observed values fall in the lower or upper 5^{th} percentile.

Comparing the results from the SNM and EXP simulated datasets, it can be seen that the observed data yielded many more significant results when compared to the simulated data generated under the EXP1 model than the simulated data generated under the SNM. All of the significant values of Tajima's *D* that are significant under the EXP1 model but not under the SNM are at the positive end of the distribution. These are MRPS05 in African *Bos taurus* and *Bos indicus*, MRPS14 in African *Bos taurus*, PRF1 in African *Bos taurus*, TLR4 in European *Bos taurus*, TLR5 in *Bos indicus* and TRIF in *Bos indicus*. It appears that the number of high positive values that was seen in the sequenced data would not be expected under the neutral EXP1 model. The 95% confidence interval range is shifted towards the negative end under the EXP1.

The pattern of significant values in the analyses of Fay and Wu's *H* is slightly different. All of the values that are significant under the SNM are also significant under the EXP1. Eight additional values are significant under the EXP1, five of which are at the positive

end of the distribution and three at the negative end of the distribution. The range of the 95% confidence interval is typically wider under the SNM than under the EXP leading to fewer significant values under the SNM. In some cases there is no difference between the two models.

Table 4.10 Probability of observed Tajima's *D* and Fay and Wu's *H* and confidence intervals determined by coalescence analysis using MS. Table continues on following page

(a)

Gene	Population	Observed *D*	95% CI *D* SNM	Observed *H*	95% CI *H* SNM
ART4	Africa	0.721	-1.888 to 1.901	0.779	-3.268 to 1.263
ART4	Europe	-0.146	-1.743 to 1.857	0.397	-3.290 to 1.178
ART4	India	0.332	-1.640 to 1.767	-2.338	-11.775 to 5.723
CD2	Africa	0.051	-1.749 to 1.786	-1.200	-5.621 to 2.458
CD2	Europe	-0.576	-1.702 to 1.600	-9.511	-10.029 to 4.708
CD2	India	-0.236	-1.725 to 1.743	-7.775	-10.814 to 4.987
FEZL	Africa	-0.191	-1.508 to 1.949	0.366	-1.673 to 0.523
FEZL	Europe	-1.244	-1.772 to 1.937	0.898	-4.013 to 1.477
FEZL	India	-1.018	-1.817 to 1.722	1.784	-5.169 to 2.070
IL2	Africa	-0.086	-1.164 to 1.565	-1.253	-1.253 to 0.263
IL2	Europe	**-1.754**	**-1.739 to 1.791**	**-5.670**	**-4.629 to 2.114**
IL2	India	-0.769	-1.728 to 1.832	0.156	-3.662 to 1.437
IL13	Africa	**-2.156**	**-1.768 to 1.852**	1.621	-5.764 to 2.779
IL13	Europe	-0.998	-1.750 to 1.798	1.756	-5.951 to 2.508
IL13	India	0.741	-1.703 to 1.926	2.130	-6.563 to 2.589
MRPL30	Africa	-0.248	-1.451 to 1.758		-1.515 to 0.545
MRPL30	Europe	0.264	-1.676 to 1.838		-3.703 to 1.625
MRPL30	India	-0.848	-1.743 to 1.890		-3.056 to 1.437
MRPS05	Africa	1.363	-1.823 to 1.781		-4.497 to 2.209
MRPS05	Europe	-0.362	-1.591 to 1.897		-6.344 to 3.013
MRPS05	India	0.628	-1.732 to 1.747		-11.567 to 5.333
MRPS14	Africa	1.542	-1.515 to 1.918		-1.593 to 0.502
MRPS14	Europe	-0.307	-1.124 to 1.679		-1.477 to 0.256
MRPS14	India	-0.175	-1.162 to 1.567		-1.316 to 0.260
PRF1	Africa	-1.562	-1.562 to 1.831	**-4.800**	**-2.267 to 0.800**
PRF1	Europe	0.820	-1.112 to 1.464	0.267	-1.600 to 0.267
PRF1	India	1.604	-1.562 to 1.831	0.622	-1.867 to 0.800
TLR2	Africa	-0.507	-1.562 to 1.831	0.711	-2.844 to 0.800
TLR2	Europe	-1.094	-1.755 to 1.720	**-4.000**	**-3.911 to 1.867**
TLR2	India	0.454	-1.687 to 1.677	0.800	-4.311 to 2.133
TLR4	Africa	-0.184	-1.401 to 1.743	0.444	-1.422 to 0.533
TLR4	Europe	1.439	-1.401 to 1.743	-0.622	-1.422 to 0.533
TLR4	India	-0.108	-1.720 to 1.726	-6.311	-6.978 to 3.289
TLR5	India	1.303	-1.112 to 1.464	0.178	-1.333 to 0.267
TLR9	Africa	1.753	-1.132 to 1.753	0.000	-1.333 to 0.533
TLR9	Europe	**-1.055**	**-1.055 to 1.444**	**-1.500**	**-1.500 to 0.286**
TLR9	India	0.820	-1.112 to 1.464	**0.267**	**-1.600 to 0.267**
TNFAIP8L1	Africa	-0.769	-1.513 to 1.858	0.337	-1.705 to 0.505
TNFAIP8L1	Europe	-0.616	-1.749 to 1.871	-0.838	-4.778 to 1.7904
TNFAIP8L1	India	0.096	-1.762 to 1.840	-0.987	-3.619 to 1.247
TRIF	Africa	-1.283	-1.631 to 1.838	**-19.179**	**-8.763 to 4.337**
TRIF	Europe	0.073	-1.695 to 1.790	-7.137	-13.725 to 6.110
TRIF	India	1.027	-1.673 to 1.725	-5.576	-8.589 to 3.844
TYROBP	Africa	-0.137	-1.781 to 1.932	**-3.779**	**-3.679 to 1.627**
TYROBP	Europe	-1.607	-1.731 to 1.986	**-7.356**	**-7.251 to 3.251**
TYROBP	India	0.745	-1.718 to 1.749	1.887	-5.239 to 2.286

(b)

Gene	Population	Observed *D*	95% CI *D* EXP1	Observed *H*	95% CI *H* EXP1
ART4	Africa	0.721	-2.056 to 1.202	0.779	-3.079 to 1.032
ART4	Europe	-0.146	-2.103 to 0.896	0.3968	-1.775 to 0.906
ART4	India	0.332	-2.271 to 0.787	-2.338	-9.861 to 4.537
CD2	Africa	0.051	-2.174 to 1.129	-1.200	-4.800 to 2.032
CD2	Europe	-0.576	-2.357 to 0.560	**-9.511**	**-7.849 to 3.207**
CD2	India	-0.236	-2.225 to 0.843	-7.775	-9.411 to 4.026
FEZL	Africa	-0.191	-1.508 to 1.523	0.366	-1.529 to 0.471
FEZL	Europe	-1.244	-2.165 to 0.941	0.898	-2.483 to 1.127
FEZL	India	-1.018	-2.219 to 0.938	**1.784**	**-4.182 to 1.662**
IL2	Africa	-0.086	-1.164 to 1.531	-1.253	-1.516 to 0.263
IL2	Europe	-1.754	-2.305 to 0.728	**-5.670**	**-3.281 to 1.546**
IL2	India	-0.769	-2.140 to 1.143	0.156	-2.494 to 1.152
IL13	Africa	-2.156	-2.225 to 0.945	1.621	-5.984 to 2.247
IL13	Europe	-0.998	-2.316 to 0.524	1.756	-4.816 to 1.787
IL13	India	0.741	-2.238 to 1.00	**2.130**	**-4.502 to 2.095**
MRPL30	Africa	-0.248	-1.451 to 1.557		-1.515 to 0.515
MRPL30	Europe	0.264	-2.226 to 0.617		-2.952 to 1.214
MRPL30	India	-0.848	-2.140 to 1.0275		-3.091 to 1.160
MRPS05	Africa	**1.363**	**-2.143 to 0.830**		-4.288 to 1.830
MRPS05	Europe	-0.362	-2.378 to 0.638		-5.212 to 2.114
MRPS05	India	0.628	-2.251 to 0.724		-9.177 to 4.242
MRPS14	Africa	**1.542**	**-1.515 to 1.488**		-1.732 to 0.450
MRPS14	Europe	-0.307	-1.124 to 1.636		-1.221 to 0.246
MRPS14	India	-0.175	-1.162 to 1.554		-1.316 to 0.260
PRF1	Africa	**-1.562**	**-1.562 to 1.416**	**-4.800**	**-1.956 to 0.711**
PRF1	Europe	0.820	-1.111 to 1.464	**0.267**	**-1.067 to 0.267**
PRF1	India	**1.604**	-1.562 to 1.491	0.622	-1.956 to 0.711
TLR2	Africa	-0.507	-1.562 to 1.830	**0.711**	**-1.867 to 0.711**
TLR2	Europe	-1.094	-1.873 to 1.178	**-4.000**	**-3.822 to 1.778**
TLR2	India	0.454	-1.901 to 1.280	0.800	-3.689 to 2.044
TLR4	Africa	-0.184	-1.401 to 1.641	0.444	-1.422 to 0.533
TLR4	Europe	1.439	-1.401 to 1.641	-0.622	-1.422 to 0.533
TLR4	India	-0.108	-2.000 to 1.233	**-6.311**	**-5.867 to 3.200**
TLR5	India	1.303	-1.112 to 1.463	0.178	-1.600 to 0.267
TLR9	Africa	**1.753**	**-1.132 to 1.753**	0.000	-1.067 to 0.533
TLR9	Europe	**-1.055**	**-1.055 to 1.444**	**-1.500**	**-1.500 to 0.286**
TLR9	India	0.820	-1.112 to 1.464	**0.267**	**-1.600 to 0.267**
TNFAIP8L1	Africa	-0.769	-1.513 to 1.661	0.337	-1.421 to 0.474
TNFAIP8L1	Europe	-0.616	-2.279 to 0.663	-0.838	-3.360 to 1.287
TNFAIP8L1	India	0.096	-2.072 to 1.152	-0.987	-2.563 to 1.022
TRIF	Africa	-1.283	-2.246 to 0.987	**-19.179**	**-8.179 to 3.511**
TRIF	Europe	0.073	-2.362 to 0.636	-7.137	-10.348 to 4.240
TRIF	India	**1.027**	**-2.302 to 0.817**	-5.576	-6.918 to 3.169
TYROBP	Africa	-0.137	-2.175 to 0.885	**-3.779**	**-3.458 to 1.389**
TYROBP	Europe	-1.607	-2.348 to 0.679	**-7.356**	**-5.864 to 2.359**
TYROBP	India	0.745	-2.255 to 1.009	**1.888**	**-4.234 to 1.887**

Simulated populations were generated under a) SNM and b) EXP1 model. The probability is the probability of obtaining the same or lower value as the observed value in the simulated data. The upper and lower 5th percentiles are indicated in red, and values falling outside the 95% confidence interval are in bold.

4.3.7.2 SCANMS analyses (sliding window analyses)

Tajima's *D*, Fu and Li's *D** and Fu and Li's *F** were calculated in sliding windows across each gene. A window size of 150 bp and a slide of 25bp were used in all analyses. To determine the significance of these values, DnaSP calculates the value for each window separately without taking into account that multiple tests are being performed. In order to overcome this, SCANMS (Ardell 2004) generates an empirical distribution of minimum and maximum values observed in each replicate simulated population. The simulated populations are generated using MS (Hudson 2002). The empirical distribution can then be used to determine the significance of the observed values.

As in the MS analysis (section 4.3.7.1), MS was used to generate 1000 simulated populations under the SNM and the EXP. The observed values of Tajima's *D*, Fu and Li's *D** and Fu and Li's *F** were compared to the empirical distributions generated from the simulated data (Table 4.11).

No significant windows were identified in MRPL30, TLR2, TRIF, TYROBP.

For this analysis, two levels of significance were considered. Table 4.11 shows all those sliding windows for which the observed value was in the first percentile (or the 99th percentile) of the empirical distribution, i.e. the observed value was equal to or lower (or higher) than 990 out of the 1000 replicates. Those windows marked with an asterisk had an observed value that was equal to or lower (or higher) than the lowest (or highest) value obtained from the simulations.

Table 4.11 Windows identified as having significant values of Tajima's *D* (D), Fu and Li's *F* (F*)* and Fu and Li's *D* (D*)* by coalescent simulation under a) SNM and b) EXP1 model. Table continues on following pages.

(a) SNM

Gene	Population	*D* (SNM)	*F** (SNM)	*D** (SNM)
ART4	Europe			4880-5104 (4)
ART4	India			9407-9557 9432-9607* (2) 9482-9632
CD2	Africa			7099-7248 7124-7348* (4) 7224-7373
FEZL	Europe		388-612 (4)	388-612 (4)
IL2	Africa			6613-6887* (6)
IL2	India	6638-6837 (3)	6638-6837 (3)	6638-6837 (3)
IL13	Africa	1473-1622	1473-1622	1473-1622
MRPS05	Europe			10476-10726 (5)
MRPS05	India			
MRPS14	Europe			8104-8278* (2)
MRPS14	India			8104-8328* (4)
PRF1	Africa	4756-4980 (4)	4756-4980 (4)	4756-4980 (4)
PRF1	Europe			4231to 4455* (4)
PRF1	India			4481-4705 (4)
TLR4	Europe			
TLR5	India			2181-2455*(6)
TLR9	Africa	4495-4719* (4)	4495-4719* (4)	4495-4719* (4)
TLR9	Europe	4445-4719* (6)	4445-4719* (6)	4445-4719* (6)
TLR9	India			4445-4719* (6)
TRIF	Africa			
TRIF	Europe			
TRIF	India			

(b)EXP1

Gene	Population	*D* (EXP1)	*F** (EXP)	*D** (EXP)
ART4	Europe			4880-5104 (4)
ART4	India			9407-9632* (4)
CD2	Africa			7074-7223 7099-7373* (6) 7249-11344
FEZL	Europe			
IL2	Africa			6613-6887* (6)
IL2	India			
IL13	Africa			
MRPS05	Europe			10476-10701* (4) 10576-10726
MRPS05	India			10476-10751 (6)
MRPS14	Europe			8104-8278* (2)
MRPS14	India			8104-8328* (4)
PRF1	Africa			
PRF1	Europe			4231to 4455* (4)
PRF1	India			4481-4705 (4)
TLR4	Europe	10182-11951		
TLR5	India			2181-2455*(6)
TLR9	Africa	4445-4619 (2) 4495-4719* (4) 4595-4769 (2)	4495-4719* (4)	4495-4719* (4)
TLR9	Europe	4445-4719* (6)	4445-4719* (6)	4445-4719* (6)
TLR9	India			4445-4719* (6)
TRIF	Africa			1585-1809 (4) 2785-3034 (4)
TRIF	Europe			2810-3034 (4)
TRIF	India			1710-1934 (4)

Sliding window analysis was performed with a window size of 150bp and a slide of 25bp. Significant windows were those windows with an observed value that fell in the lower or upper first percentile of the simulated distributions of maximum and minimum values for each statistic for each gene. Windows marked with asterisk had an observed value that was more extreme than any simulated values. When a window positions is followed by a number in parentheses, this indicates the number of overlapping 150bp windows that were found to be significant at that level.

4.4 Discussion

Resistance against disease provides an obvious selective advantage through increased survival and fitness. It is not surprising, therefore, that loci involved in the immune response, have previously been identified as candidates for selection studies (Akey et al. 2004; Jiggins and Kim 2007; Schlenke and Begun 2003; Schlenke and Begun 2005). Given the change of population structure associated with domestication and the novel disease challenges likely to accompany these changes, immune response genes are even more likely to be important in the recent adaptation of domesticated species, such as cattle, as disease resistance is associated with increased productivity. Genes known or thought to be involved in susceptibility to disease therefore make strong candidates to look for signs of recent selection.

Of the 15 candidate genes analysed here, there was variability both between genes and between populations in terms of the strength of evidence for selection. IL2 in European *Bos taurus* and IL13 in African *Bos taurus* both showed statistically significant deviations from neutrality at multiple tests of neutrality. These are considered to be the most likely candidates for selection emerging from this study. Other genes, notably CD2 and TRIF, did not have many significant results showing deviations from neutrality but have emerged as interesting candidates for other reasons, including non-synonymous substitutions, haplotype diversity and linkage disequilibrium (LD). These results are discussed in more detail below.

In most cases, significant results are population-specific. Very few results were significant across all populations for the same gene. This pattern of results is consistent with specific local adaptations, as would be expected at disease-resistance loci. The same pattern has also been reported in a study of human genes, where it was also interpreted as a signal of local selection (Akey et al. 2004). In further support of local adaptations, different genes have significant results in different populations, making it unlikely that a single demographic feature of one population was responsible for deviations from the neutral expectation. Due to the number of tests involved, individual significant results should be treated with caution. However, when the deviation from neutrality is strong and/or consistent across a number of tests, it could still indicate a result of interest.

Fu and Li recommend the use of the tests requiring an outgroup (D and F) over the tests that do not require an outgroup (D* and F*). Furthermore, they recommend that this outgroup be from a closely related population or species so that recurrent or back

mutations are less likely. Plains Bison is probably the most closely related species to cattle (followed by Gaur) (Buntjer et al. 2002; Finlay 2006), but these may still be distant enough that parallel mutations and back mutations might have occurred. It is therefore prudent to consider F^* and D^*, which do not use an outgroup, in addition to F and D, which do. Furthermore, the sliding window analysis carried out using SCANMS could only be used to analyse D^* and F^*.

4.4.1 Interleukins

The most robust signatures of non-neutral evolution came from the two interleukin (IL) genes included in the study. Interleukins are a family of cytokines originally identified as playing a role in signalling between white blood cells (leukocytes). Leukocytes are major players in the immune response to infectious agents. IL2 and IL13 are central regulators of the immune response. Together, they regulate the synthesis of interferon-gamma (Minty et al. 1993), which itself is an important immunoregulatory cytokine with antiviral properties (Schroder et al. 2004). As important signalling molecules, IL2 and IL13 are well placed to affect the gene expression of a host of other molecules in response to various pathogenic challenges.

4.4.1.1 IL2

Of the 14 polymorphic sites identified in IL2 in cattle, only a single site was identified in the coding sequence. This site is synonymous and was found to be polymorphic only in *Bos indicus*.

Tajima's D, Fu and Li's D^* and F^*, Fay and Wu's H and Fu's Fs were all found to be significantly negative in European *Bos taurus*. Taken together, these results are highly suggestive that positive selection has acted on the IL2 locus in the European population. In contrast, the only departure from neutral expectation in either the African *Bos taurus* or *Bos indicus* was a negative Fay and Wu's H result for African *Bos taurus*, which was close to significant. Only one polymorphic site was identified in African *Bos taurus*, and this site showed a high frequency of the derived allele (as determined from the Plains Bison outgroup), resulting in the negative value of Fay and Wu's test. The SCANMS sliding-window analysis identified a number of windows that had a significant value of Fu and Li's D^* in African *Bos taurus* but these results must be taken cautiously as the values of the statistics were calculated from a single polymorphic site.

The negative values of the neutrality tests for European *Bos taurus* are indicative of an excess of recently-derived, low frequency variants, consistent with a selective sweep at

this locus. When the significance of the Tajima's *D* deviation was assessed based on simulated populations created using a coalescence-based approach, it was no longer significant when the populations were simulated under the EXP1 model, which may be indicative of an expanding European *Bos taurus* population. Significance was observed under the standard neutral model (SNM), which maintains a constant population size. However, Fay and Wu's *H* remained significant under both models, suggesting that demographics alone are unlikely to be responsible for the observations. A deviation from neutrality due to demographic history is also likely to affect all loci in a similar fashion, which is not observed in the data. Instead, IL2 appears to be a genuine candidate for selection in the European *Bos taurus* population.

4.4.1.2 IL13

Like IL2, only one SNP was identified in the coding region of IL13 in cattle. This was a non-synonymous SNP that was polymorphic in *Bos indicus*, and African *Bos taurus*. A single minor allele was identified in the latter. An additional site in the coding region was identified where the G nucleotide in cattle was an A nucleotide in Plains Bison. This site was not found to be polymorphic in cattle.

Although there was significantly high haplotype diversity in the *Bos indicus* population, no deviations from neutrality were detected by any of the neutrality tests. The African *Bos taurus* population, however, showed strong evidence for non-neutral evolution as indicated by the significant and highly significant test results. The most negative value of Tajima's *D* across all genes and all populations was -2.16 ($p<0.001$) for IL13 in the African taurine population. Fu and Li's *D**, *F**, *D* and *F* were also significantly negative in African *Bos taurus*. All these results support a selective sweep in the African taurine population. Fu and Li's *D*, *D** and *F** were approaching significance in European *Bos taurus*, suggesting that the selective pressure might be common to both populations. Alternatively this selective pressure may have occurred in the ancestral *Bos taurus* population. Although Fu and Li's statistics are measuring the relative excess or dearth of recent mutations, the ancestral *Bos taurus* population divergence relatively recently on an evolutionary timescale. Fu's *Fs*, which was highly significant in African *Bos taurus* but also significant in European *Bos taurus,* also suggests an excess of recent mutations. The haplotype network for IL13 further supports this view (Figure 4.5), with both African- and European-specific low frequency haplotypes radiating from common shared haplotypes.

IL13 is quite unusual in that it showed a significantly *positive* deviation from neutrality for Fay and Wu's *H* in African *Bos taurus*. This seems somewhat contradictory to the significantly negative value of Fu and Li's *F*, *D*, *F**, *D**. Fay and Wu's *H* yielded more significant values than any of the other tests, suggesting that it may be more sensitive to deviations from the assumptions of the model. Given the large number of tests involved, it is expected that some of the significant results are observed are due to chance (*i.e.* a Type I error). The marginal significance of the positive Fay and Wu's *H* in IL13 and its contradiction by a number of other tests suggest that it may be one such result and should therefore be treated with caution.

4.4.2 CD2

CD2 was reported to have been influenced by positive selection in the bovine lineage as evidenced by an elevated omega value (1.41), which is the ratio of non-synonymous to synonymous substitutions, d_N/d_S (Lynn et al. 2005). This effect was found to be more pronounced in the extracellular domain of the protein where an omega value of 2.56 was observed. Furthermore, a computational approach identified 20 amino acid sites predicted to be under positive selection, 18 of which were in the extracellular domain of the protein (Lynn et al. 2005).

In the resequencing study presented here, no significant deviations from neutrality were identified in CD2 at the whole gene level in any of the populations tested. The value of Fay and Wu's *H* in European *Bos taurus* was approaching significance, suggesting that there is an excess of high frequency derived alleles in this population. This result was found to be significant when assessed by coalescence-based analysis with the EXP1 model, but only marginally significant when the SNM was used.

Sliding window analysis, using SCANMS, identified a specific region that showed significantly positive values of Fu and Li's *D** in African *Bos taurus* under both the SNM and the EXP1 model. This region, from 7074 to 7373bp (assembly positions) spans exon 3. Exon 3 is remarkable in that eight non-synonymous polymorphisms but only a single synonymous polymorphism were identified. Figure 4.7(b) (sliding window) shows that π is especially high in this region and in fact, this was among the highest π values observed. Tajima's D was negative in African *Bos taurus*, but positive in *Bos indicus* and European *Bos taurus*. Inspection of this region in Figure 4.4 suggests that there are a number of divergent haplotypes at high frequencies in both populations. This pattern is not observed in the rest of the gene, but seems specific to exon 3. The region encoded by exon three contains the Ig-like C2 set domain of CD2. This domain resembles an Ig constant

domain, and is characteristic of T-cell surface antigens CD2, CD4, and CD80, as well and VCAM (vascular cell adhesion molecules) and ICAM (intercellular cell adhesion molecules) (Tangye et al. 2000). In addition to important cell-adhesion and immune response roles, these antigens are known to be important targets for pathogens.

CD2 is similar to an open reading frame found in African Swine Fever Virus (ASFV) (Rodriguez et al. 1993) that has been implicated in immune suppression during infection (Borca et al. 1994). It has previously been suggested that exploitation by viruses might have placed CD2 under selective pressure (Lynn et al. 2005) as the low affinity of CD2 for its counter receptors might be easily exploited by viruses (Wang 2002). Although speculative, it seems reasonable to suggest that CD2, and specifically exon 3, may be under selection due to novel viral exposure following domestication.

4.4.3 TRIF

The sequence of TRIF is highly polymorphic with thirty-three SNPs in the coding sequence, seventeen of which are non-synonymous. Haplotype diversity was low in all populations, and this was found to be significant. In addition, an unusually high number of pairs of sites showed significant LD at the $p<0.05$ level (Figure 4.6). Fay and Wu's H was highly significant in the African *Bos taurus* ($H = -19.17$). This was the most significant result for Fay and Wu's H for all populations and all genes ($p = 0.001$). A negative value of Fay and Wu's H is indicative of an excess of high frequency derived alleles. π is especially high in European *Bos taurus*.

Within the *Bos taurus* populations, highly divergent haplotypes can be seen. Several of these highly divergent haplotypes reach appreciable frequency in the populations. This is evident in the network of TRIF haplotypes (Figure 4.5) and the alignment of polymorphic sites (**Figure 4.4**). This pattern was not observed at the other loci sequenced in the same population, with the exception of CD2, which also shows unusually balanced haplotypes, particularly in exon 3 as discussed above. Within the European *Bos taurus*, there is some distinction between northern *Bos taurus* and Southern/Mediterranean *Bos taurus* as is evident from Figure 4.4. Southern breeds, i.e. Romagnola, Sikias, Anatolian Black, Mertolenga and Alentejana are more frequency carriers of the divergent haplotype.

4.4.4 PRF1 and Toll-like Receptors (TLRs)

PRF1, TLR2, TLR4, TLR5, TLR7 and TLR9 were sequenced in the smaller screening panel, comprised of 15 animals (5 European *Bos taurus*, 5 African *Bos taurus* and 5 *Bos*

indicus) and one outgroup (Gaur). For each of these genes, a single PCR fragment was sequenced, yielding between 412bp (TLR5) and 1114bp (TLR2) with an average of 916bp.

4.4.4.1 PRF1

No significant deviations from neutrality were detected in the *Bos indicus* or European *Bos taurus* populations using the basic neutrality tests. The positive value of Tajima's D observed in the *Bos indicus* population was found to be significant under the EXP1 model but not under the SNM. The SCANMS-based sliding window analysis identified a region from 4231-4455 which had a significant D* value in European *Bos taurus*, and a region from 4481-4705 which had a marginally significant value in *Bos indicus*. Because of the large number of tests made and lack of supporting results, these observations of deviations from neutrality in *Bos indicus* and European *Bos taurus* can not be considered robust.

In contrast, PRF1 in African *Bos taurus* showed a number of significant deviations from neutrality. Tajima's *D* was significantly negative in African *Bos taurus*, but positive in both the other populations. Fu and Li's *D** and *F** are also significantly negative, but Fu and Li's *D* and *F* are positive, with only *D* being marginally significant. Fay and Wu's *H* was also significantly negative in the African *Bos taurus*. When the significance of Tajima's *D* and Fu and Li's *H* were assessed by coalescent simulation using MS, the value of Tajima's *D* in Africa was significant under the EXP1 model but only marginally significant under the SNM. Conversely, when the sliding window analysis was carried out using SCANMS, windows having significant values of Tajima's *D*, Fu and Li's *D** and Fu and Li's *F** were identified in the African *Bos taurus* population when the SNM was applied, but no significant windows were identified when the EXP1 model was applied.

This number of significant results is suggestive of positive selection on the PRF1 gene in African *Bos taurus*. However, inspection of the alignment (Figure 4.4) reveals a more likely explanation. A single Somba individual (SOM60) contains all the polymorphism in the African *Bos taurus* sample. The haplotype SOM60-A carries a C allele at position 396, a C allele at 745 and a G allele at 812. These three alleles are not found in any other *Bos taurus* animal but were present at frequencies of 100%, 100% and 70% in the *Bos indicus* sample. This is strongly suggestive of a *Bos indicus* introgression into the Somba population. This introgression is also apparent in the network diagram of PRF1 haplotypes (Figure 4.5). PRF1 is not the only gene showing evidence of *Bos indicus*

introgression in *Bos taurus*, but perhaps because of the small sample size and small number of polymorphic sites, this demographic effect is the dominant signal in the data.

Despite the small number of polymorphic sites, it is apparent that there is an exceptional divergence between *Bos taurus* and *Bos indicus* at this locus. The sliding window analysis of divergence (Figure 4.7) shows that the divergence between African and European *Bos taurus* populations is extremely low, whereas the divergence between *Bos indicus* and either *Bos taurus* population is among the highest observed across all the genes sequenced, despite the few polymorphic sites and limited sample size. Four non-synonymous polymorphisms were identified in cattle, but only a single synonymous polymorphism was observed. The non-synonymous polymorphism appears to be fixed between *Bos indicus* and *Bos taurus* (excluding the instance of *Bos indicus* introgression mentioned above), as does a non-coding SNP. Three additional non-synonymous substitutions and a single non-coding substitution separated the Gaur and the cattle sequences.

4.4.4.2 TLR2

TLR2 also shows a high proportion of non-synonymous polymorphisms compared to synonymous ones (10/13). Unlike PRF1, however, this does not seem to be due to divergence between *Bos taurus* and *Bos indicus*. Five polymorphisms are shared by both species (*i.e.* both alleles are found in taurine and indicine animals) and none of them can be used to differentiate one population or species from another.

When the significance of Tajima's D and Fay and Wu's H were assessed by coalescent simulation, Fay and Wu's H was found to be significantly negative in both taurine populations under the EXP1 model but only marginally significant in the African population when the SNM was used. Fu & Li's D and D^* are significantly positive in *Bos indicus*, though this is not reflected in the other tests and so should again be interpreted with caution.

4.4.4.3 TLR4

TLR4 showed limited polymorphism in European and African *Bos taurus* (two sites which were polymorphic in both populations). This was not reflected by *Bos indicus*, which was found to have fifteen polymorphic sites (including one of the sites that was polymorphic in *Bos taurus*). These results are similar to those observed by White et al. (2003) who identified 29 polymorphic sites in a *Bos indicus* population (n=13) and only 5 in a larger *Bos taurus* population (n=26). TLR4 had the highest diversity (θ (per site) =

0.00482, μ (per site) = 0.00471). Fay and Wu's *H* was found to be significantly negative in the *Bos indicus* population when assessed by coalescence-based analysis using the EXP1 model, and marginally significant when the SNM was used. None of the other tests showed any significant deviations from neutrality. The results presented here are weakly suggestive of non-neutral patterns of polymorphism in the *Bos indicus* population. A pattern of polymorphism indicative of balancing selection has previously been reported in a *Bos indicus* population (White et al. 2003). Intriguingly, purifying selection at TLR4 in a human population has also been reported (Smirnova et al. 2001) When the data presented here are taken together with the results of White et al. (2003), they suggest that there is not enough power to detect a signal of selection in the data set of limited sample size and sequence coverage.

4.4.4.4 TLR9

Only three polymorphic sites were identified in TLR9 in the cattle population screened, all of which were non-synonymous. Only three African *Bos taurus* animals were successfully sequenced, which obviously limits the power of any tests in this population. Nevertheless, Tajima's *D* is significantly positive in African *Bos taurus*. This value was found to significant by coalescence-based analysis using the EXP1 model and was close to significant when the SNM was used. In addition, under these conditions, Fay and Wu's *H* and Tajima's *D* were significantly negative in Europe, indicating an excess of low frequency alleles and an excess of high frequency derived alleles respectively. These values are due to a single polymorphic site at position 4580 with one count of the ancestral allele as determined by reference to the outgroup sequence (see Figure 4.4). This is the only polymorphic site in the European *Bos taurus* population.

4.5 Conclusions

Resequencing permits a more in depth analysis of candidate genes than microsatellite studies. The majority of the genes analysed did not show strong evidence of selection. This is not surprising, as however carefully candidate loci are chosen it is unlikely that many of the them will actually have been subject to selection of sufficient strength and/or recently enough to be detected by neutrality tests. This is particularly true when the populations sampled are quite small, as here. Not only is the lack of significant deviations from neutrality in several genes unsurprising, it is also reassuring. Neutrality tests can be systematically biased by a demographic history or population sub-structure that violates the assumptions of the underlying models. Given the variety of test results observed for different genes, it does not seem likely that such systematic biases are responsible. Cattle demographic history is investigated further in the next chapter.

The strongest candidates for having been affected by positive selection are the interleukins IL2 and IL13. Both of these genes show strong evidence for selection in one but not all of the populations examined: IL2 in European *Bos taurus* and IL13 in African *Bos taurus* (Table 4.12). These candidates were both selected based on results of an inter-species PAML analysis (Freeman et al. In Prep.; Lynn et al. 2005), which indicated that positive selection occurred on the bovine lineage. The observation here of non-neutral evolution in a single population suggests that the historical selective pressures that originally gave rise to the signal of positive selection in the inter-species PAML analyses are no longer driving evolution equally across modern populations.

Table 4.12 Summary of neutrality test results for IL2 and IL13.

The values of Tajima's D, Fu and Li's D, Fu and Li's F, Fu and Li's D*, Fu and Li's F*, Fay and Wu's H, and Fu's Fs observed in IL2 and IL13 in the African *Bos taurus*, European *Bos taurus* and *Bos indicus* populations are given below. Significant deviations are shown in bold red print.

Gene	Pop	Tajima's D	Fu and Li's D	Fu and Li's F	Fu and Li's D*	Fu and Li's F*	Fay and Wu's H	Fu's Fs
IL2	Africa	-0.086	0.627	0.512	0.65	0.52	-1.253	0.381
	Europe	**-1.754**	-1.144	-1.613	**-2.487**	**-2.647**	**-5.68**	**-3.667**
	India	-0.7689	-1.479	-1.507	-1.278	-1.311	0.156	-2.008
IL13	Africa	**-2.156**	**-3.521**	**-3.731**	**-2.916**	**-3.129**	1.621	**-3.877**
	Europe	-0.998	-2.034	-2.02	-1.788	-1.804	1.756	**-4.53**
	India	0.741	1.232	1.296	1.131	1.182	2.13	-2.505

There are also some suggestions of non-neutral evidence even where neutrality tests fail to reject the null hypothesis. Such inferences are possible due to the extra information that resequencing provides. The T-cell surface receptor CD2, for example, shows abnormally high levels of nucleotide diversity and particularly high numbers of non-synonymous SNPs in the third exon, despite a lack of significant results for the whole gene. This exon encodes an Ig-like domain, which are important for protein interactions. Other T-cell surface receptors that share this domain are known to be important targets for pathogens, making this result particularly interesting.

The final gene of interest, TRIF, was selected as a candidate gene based on its position in a region of low microsatellite diversity found close to a QTL for trypanotolerence (ref Park). High levels of diversity were observed at this locus, with thirty three SNPs being identified in the coding region. TRIF showed the lowest levels of haplotype diversity, given the number of polymorphic sites and unusually high levels of pairwise linkage disequilibrium (LD), particularly in African *Bos taurus*. These observations, when taken together with the QTL analysis and the known function of TRIF in the immune response to pathogens, make TRIF an extremely interesting candidate for a key role in trypanotolerence.

Despite the small region sequenced and the limited sample size, some trends suggestive of non-neutral evolution have also emerged from the smaller screening panel. The hypothesis of neutrality at PRF1 in the African *Bos taurus* population was rejected by several tests: Tajima's's D, Fu and Li's F*, Fu and Li's D* and Fay and Wu's H. TLR9 also shows evidence of non neutral evolution as indicated by a significantly positive value

of Tajima's D in the African *Bos taurus* population. These results must be interpreted with caution, however. As can be seen in PRF1, demographic effects, such as the introgression of a single *Bos indicus* haplotype, have a greater impact when the sample size is small; in TLR9, a single polymorphic site produced a significant value. In such a case, it must be asked whether the pattern of polymorphism that is deviating from neutral expectation is more likely to be explained by some demographic factor, such as migration/introgression in PRF1, or an artefact of the methodology; Fay and Wu's H in TLR9 might be explained by a recurrent or back mutation at position 4580 in one of the species such that the supposedly derived allele is not in fact the derived allele.

These examples, combined with the findings of White et al. (2003) in TLR4, reinforce the need for larger sample sizes in this kind of analysis. Nonetheless, the interesting results obtained for IL2, IL13, CD2 and TRIF demonstrate the potential value of candidate gene resequencing in general and, the selection of the specific candidate genes used in this study.

5 Coalescent Simulation and Bayesian Analyses of Demographic Histories

5.1 Introduction

5.1.1 Neutrality tests and demographic history

Natural populations rarely conform to the assumptions inherent in the neutral models which form the basis of many tests and methodologies in population genetics. This includes neutrality tests based on θ, such as those described in Chapter 4, and coalescent theory, which is based on a Wright-Fisher model. One of these assumptions concerns population size. Often it is necessary to assume that the population size is either infinitely large (in the case of the Wright-Fisher model) or constant, in the case of coalescence-based methodologies. Clearly these assumptions do not always hold true.

In the case of cattle, it is commonly assumed that the species underwent a severe population bottleneck at the time of domestication. This is a reasonable assumption; undoubtedly the wild progenitor, the aurochs, was a formidable animal standing significantly taller than even the largest modern cattle at over six and half feet. Historical reports indicate that the aurochs were aggressive (Edwards et al. 2007) so domestication must have been a difficult event. The hypothesis of a severe bottleneck in the history of cattle is supported by mitochondrial DNA evidence. There is considerably less diversity in a global sample of cattle mitochondrial DNA than was observed in either impala or Grant's gazelle (Bradley and Magee 2006). These data are consistent with a limited sampling from the wild progenitor species as would be expected by a bottleneck at the time of domestication. High levels of diversity at autosomal loci, however, are not consistent with such a severe bottleneck (Vila et al. 2005). Mitochondrial DNA is inherited without recombination along the maternal lineage and is therefore only informative of a fraction of the ancestry of the individual from which the sequence was obtained. Furthermore, its smaller effective population size makes it more susceptible to demographic events and in particular to the stochastic nature of genetic drift and mutation.

5.1.2 Bayesian Skyline Plots

Skyline plots are a useful tool developed in order to indicate the most appropriate demographic model for a given data set. They are a simple representation of time from the present to the past on the x-axis and effective population size on the y-axis and are advantageous over alternative methods. Classical skyline plots (Pybus et al. 2000) bear a striking resemblance to the cityscapes after which they were named, but modifications to the method since that time have resulted in smoother, more realistic plots.

Classical skyline plots were constructed using a maximum likelihood method for estimating the demographic history of a population from a genealogy of a sample taken from that population. A major assumption of this method is that the genealogy of the sample can be reliably reconstructed. Unfortunately, this assumption does not hold true for many datasets, particularly those with lower levels of variation. This means that the high degree of phylogenetic uncertainty is not considered in the analysis. The method was refined by (Strimmer and Pybus 2001) to give the generalized skyline plot. The generalized skyline plot allows for the grouping of adjacent coalescent intervals, i.e., a coalescent event does not necessarily result in a change in population size. This method is less sensitive to phylogenetic uncertainty. This method also results in a smoother plot because the stochastic noise is removed by the grouping of small intervals. Although an improvement on the classical skyline plot, the generalized skyline plot is still dependent on the phylogenetic reconstruction of the genealogy.

The Bayesian skyline plot (Drummond et al. 2005) is a further improvement of the method and resolves the problem of uncertainty in phylogenetic reconstruction. To achieve this, the posterior distribution of the effective population size is estimated directly from the sequence data using a coalescent MCMC method. This method is more computationally intensive than either the classical or generalized skyline plots but offers several advantages over earlier methods: in addition to taking into account both the phylogenetic uncertainty and the stochastic error associated with the coalescent process, the BSP seems to be more sensitive in detecting signals of demographic history.

5.1.3 Analyses of bovid DNA using skyline plots

The demographic histories of wild and domesticated bovids have previously been investigated by using the skyline approach to analyse mitochondrial DNA sequences (Drummond et al. 2005; Finlay et al. 2007; Shapiro et al. 2004).

Using the generalized skyline plot, the population size of Beringian bison (*Bison bison*) was found to have increased until approximately 37,000 years ago, at which time the population size began to decline (Shapiro et al. 2004). This decline was found to have continued up to the present day. These results were suggested to exonerate humans as the sole cause of the declining bison population as the earliest robust evidence for significant human settlement in Beringia is at around 15,000 years before present. However, improvements to the methodology discussed above allowed the bison data of Shapiro et al to be revisited. The more sensitive BSP allowed a higher resolution plot of effective population size to be produced and resulted in a number of interesting differences. Most

notable was the rate of decline of the bison population size, which was found to increase sharply approximately 15,000 years ago. This increase occurred around the same time that significant human settlements appeared in Beringia.

A very different signal was seen in the BSPs obtained from mitochondrial DNA sequences from domesticated bovids; yak, mithan (the domesticated form of Gaur), cattle and water buffalo; and from another wild bovid, African Buffalo (Finlay et al. 2007). The data used here were isochronous (i.e. sampled at the same point in evolutionary time) and so lacked the depth of analysis that was seen in the above study as the coalescent based BSP analysis is only informative up to the point of the most recent common ancestor of the sequences used. Despite this limitation, striking results were obtained. The BSPs of yak, mithan, water buffalo and cattle were characterised by a sudden rapid increase in population size occurring within the last 10,000 years. Such a pattern was not observed in African buffalo, suggesting that it is a plausible signature of domestication associated population expansion or possibly a postglacial population recovery.

This analysis of the bison data (Drummond et al. 2005; Shapiro et al. 2004) had the advantage of ancient DNA sequences, which allow a greater depth in the analysis. Such ancient DNA data is not yet available for nuclear data on the scale presented here. Nevertheless, the analysis of isochronous sequences is worthwhile as illustrated by Finlay (2007).

5.2 Materials and Methods

5.2.1 Coalescent simulations

MS (Hudson 2002) was used to generate simulated populations under different models of demographic history (see 5.2.1.2, below). All simulations were based on an infinite-sites model, and assumed panmixis. Recombination was not included in any models.

5.2.1.1 Data

Simulated populations were generated for comparison with the observed data from 16 of the 17 sequenced regions described in Chapter 4. These were: CD2, IL2, IL13, ART4, TYROBP, TRIF, TNFAIP8L1, FEZL, MRPL30, MRPS14, MRPS05, PRF1, TLR2, TLR4, TLR5, and TLR9. TLR7 was omitted from the analysis as no polymorphic sites were identified in the region sequenced.

5.2.1.2 Demographic models

Four different models of demographic history were simulated (Figure 5.1). These are described below:

1. Standard Neutral Model (SNM): The population size was constant at 10^5 throughout the history of the population. (Figure 5.1(a)).

2. Expansion Model 1 (EXP1): The population size was constant at 10^4 up until 10,000 years ago, at which point a bottleneck occurred, reducing the population size to 10^3. From this point onwards, the population began to expand until it reached 10^5 at the present time (Figure 5.1(b)).

3. Bottleneck and Recovery Model (BTL): The population size was constant at 10^4 up until 10,000 year ago, at which point a bottleneck occurred, reducing the population size to 10^3. The bottleneck lasted for 2000 years, after which the population size increased to 10^5 at 8000 years before present and remained constant until the present time (Figure 5.1(c)).

4. Expansion Model 2 (EXP 2): The population size was constant at 10^3 up until 10,000 years ago, at which point an expansion began. The expansion continued until the population reached 10^5 at the present time (Figure 5.1(d)).

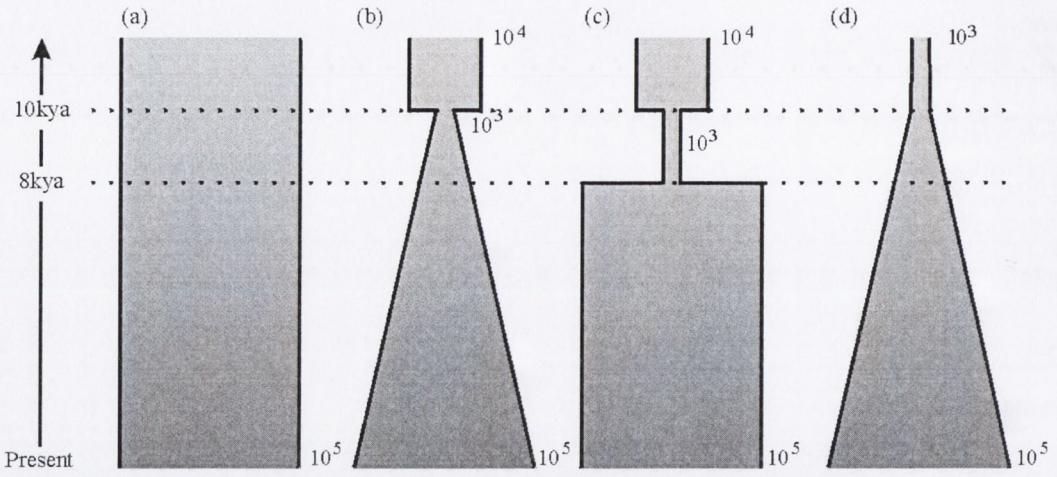


Figure 5.1 Models of demographic history simulated using MS (Hudson, 2002).

5.2.1.3 Generating simulated populations

The instructions for simulating the models are described by a single line of text, which indicates the sample size of the population to be simulated, the number of replicate populations to be simulated, the mutation parameter, θ, indicated by the switch –t and a number of optional switches, which describe how the population size has changed in the past. Additionally, the number of segregating sites can be specified by using the –s switch and the demographic history of the population can be described by a series of possible switches.

The mutation parameter, θ, is equal to $4N_0\mu$, where N_0 is the diploid population size at time 0 (i.e. at the present) and μ is the neutral mutation rate per generation across the entire locus. N_0 was set at 10^5 for all simulations. μ was determined by multiplying the mutation rate per site per generation (1×10^{-8} from (Liu et al. 2006)) by the number of sequenced sites in the locus. This results in a different value of θ for each simulated population. The number of segregating sites in the simulated populations was constrained to be the same as that observed in the sequence data.

The parameters that describe the demographic history are not affected by differences in the observed data between the different genes. As the simulations are generated using a coalescence-based approach, beginning at the present and moving backward through time until the lineages coalesce, the parameters describing the changes in population size must be specified in chronological order, beginning with the most recent and ending with the oldest change.

The following parameter settings were used to for the four models:

1. Standard Neutral Model (SNM): no demographic history is specified, resulting in constant population size. This is the default for MS.

2. Expansion Model 1 (EXP1): -G 184.207 –eN 0.025 0.1

3. Bottleneck and recovery (BTL): -eN 0.02 0.01 –eN 0.025 0.1

4. Expansion Model 2 (EXP2): -G 184.207 –eN 0.025 0.01

The –G switch is followed by a single number, α, and indicates population growth rate. The growth rate is given by $N(t) = N_0 e^{-\alpha t}$, where t is the time before present measured in units of N_0. A positive value for α indicates that the population was smaller at time t (in the past) than it is at present.

The –eN switch indicates an instantaneous change in population size and automatically sets the growth rate to zero. The –eN switch is followed by two numbers, the first indicates the time (in units of N_0) at which the change in population size occurs, and the second indicates the new population size as a proportion of N_0.

One thousand replicate simulated populations were generated for each gene and for each geographical population group (*Bos indicus*, European *Bos taurus*, African *Bos taurus*) under each of the four models.

5.2.1.4 Statistics

In total, 192 simulations (16 genes x 3 populations x 4 models) were carried out, each generating 1000 replicate populations. 'sample-stats', a program supplied with MS, was used to calculate Tajima's *D* for each of the simulated populations. The mean Tajima's *D* (Dsim) for each simulation scenario was calculated, along with the standard deviation (SDsim) of the distribution of the 1000 simulated values of Tajima's *D*. These values were used to compare the observed data (i.e. the sequencing data from Chapter 4) to the simulated data from each of the four models in order to see which gave the best fit to the observed data.

5.2.2 BEAST

BEAST (Drummond and Rambaut 2007) was used to generate Bayesian Skyline Plots by a coalescent MCMC analysis of sequence data described in Chapter 4. BEAST allows the combined analysis of multiple sequence alignments through partitioning of the input file. Parameters can be allowed to vary between partitions or can be fixed across all partitions.

5.2.2.1 Data

Sequence data from eight of the sixteen genes sequenced in Chapter 4 were combined for analysis using BEAST.

PRF1, TLR2, TLR4, TLR5, TLR7 and TLR9 were excluded from this analysis as they were sequenced in a different, smaller panel. The MRPs were excluded due to higher levels of missing data. The partitioned file was constructed such that sequence data were available for all individuals at all loci. Including the MRPs would have forced the exclusion of a number of individuals from the remaining 8 loci.

In total, 27,208bp of sequence in 37 animals was included: 3752bp for CD2, 3461bp for IL2, 3297bp for IL13, 2157bp for ART4, 3966bp for TYROBP, 4282bp for FEZL, 4130bp for TRIF and 2163bp for TNFAIP8L1. The 37 animals included 11 *Bos indicus* (22 haplotypes), and 26 *Bos taurus* (52 haplotypes), of which 17 were European *Bos taurus* (34 haplotypes) and 9 were African *Bos taurus* (18 haplotypes).

5.2.2.2 Substitution Model

Modeltest v3.7, (Posada and Crandall 1998), in conjunction with PAUP* (Swofford 2002), was used to assess which substitution model might be most appropriate for these data. Both the hLRT (hierarchical likelihood ratio test) and the AIC (Akaike Information Content) approaches were considered. Sequence data were divided into *Bos indicus* population and *Bos taurus* population groups and was tested one gene at a time.

5.2.2.3 Parameters and settings for BEAST

For each gene, haplotype alignments in NEXUS format were prepared using the application BEAUti (Bayesian Evolutionary Analysis Utility), which is included as part of the BEAST package. The BEAST input format is an XML file containing taxa information, sequence alignments and a series of blocks that describe the model and the parameters. A separate XML file was generated for each gene, and these files were combined manually to generate the partitioned file. Partitioning involves indicating individual alignments for each region in the file. Parameters can be allowed to vary between alignments, or can be fixed across alignments.

All analyses were run with a constant substitution rate of 10^{-8} changes per site per generation. The substitution rate in cattle was estimated to be 2.0×10^{-9} changes per site per year (Liu et al. 2006) although considerable rate variation was observed across chromosomal regions and functional components. Taking the overall rate, and assuming a generation time of 5 years for cattle, this gives a substitution rate of 1.0×10^{-8} changes

per site per generation. The HKY substitution model was assumed. Jeffrey's prior has been suggested to be the most appropriate prior for population size (Drummond et al. 2002) and in these analyses it was selected for the prior distribution of population size. All operators were optimised automatically.

A MCMC chain length of 10^7 was specified for the *Bos indicus* analysis and African *Bos taurus* analyses, $2x10^7$ for the European *Bos taurus* analyses and $4x10^7$ for the combined *Bos taurus* analyses. The chain length needed is dependent on the number of sequences to be analysed; the more sequences to be analysed, the greater the number of chains will be required. The actual number of chains necessary is determined by the ESS (Effective Sample Size) values of the parameters of interest; that number must be high enough to give adequate ESS values. The ESS is the number of effectively independent draws from the posterior distribution. Low ESS values (i.e. less than 100) indicate that the MCMC has not been running long enough to give a reliable estimate of the posterior distribution for that parameter. This is especially important for those parameters of particular interest. ESS values lower than 100 are flagged by Tracer (ver1.3, http://tree.bio.ed.ac.uk/software/tracer/) but the authors of BEAST suggest that this may be too liberal, and favour a lower limit of 200.

The burn-in period, the initial period of the MCMC iterative process that is discarded, was set at 10% of the chain length. Sampling for the BEAST output files, log file and tree file, was carried out every 1000 chains for those runs with 10^7 chains, every 2000 chains for those runs with $2x10^7$ chains and every 4000 chains for those runs with $4x10^7$ chains, yielding a total of 10000 samplings for each run.

Each analysis was run twice and replicate log files were combined using LogCombiner (distributed as part of the BEAST package). Tracer was used to explore the results of the analyses. Each parameter has a mean value and an ESS value. All ESS values were greater than 200, with one exception, and most were greater than 1000. The exception was the treeLikelihood4 in the combined *Bos taurus* analysis; the ESS value were below 100 and remained below 100 even when the logs were combined. The trace function of Tracer was used to examine the trace of each parameter against the generation number in order to ensure convergence to a stationary distribution. Tracer was also used to visualise the Bayesian Skyline Plots.

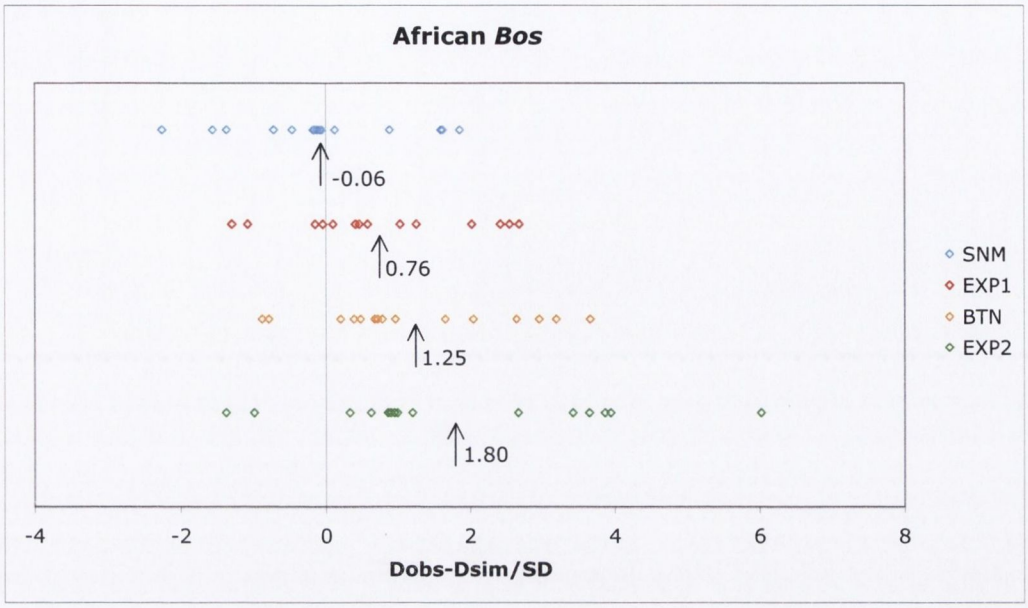
5.3 Results

5.3.1 Coalescent-based analysis (MS)

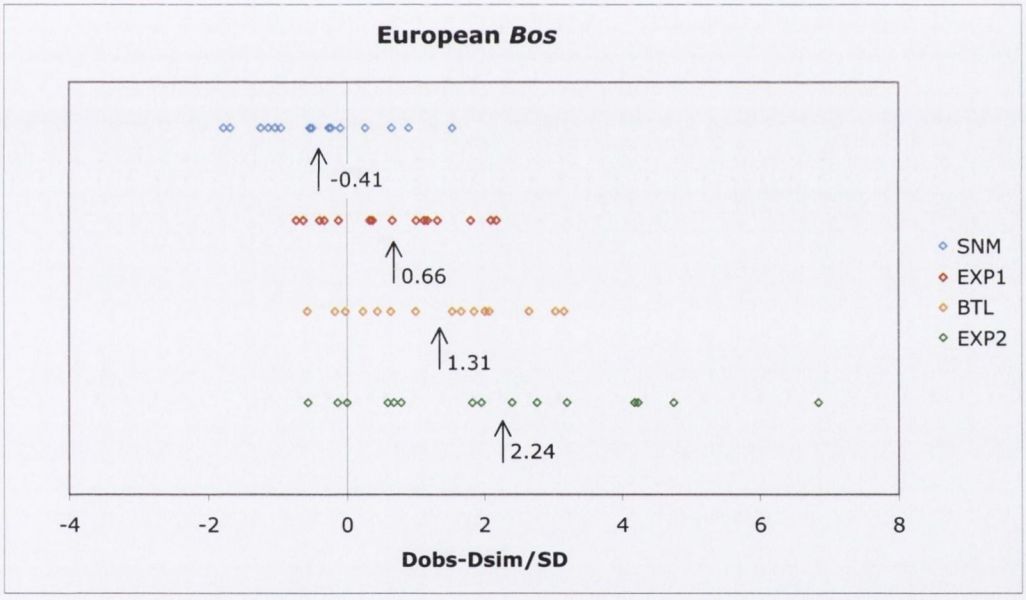
Simulated populations were generated using under four models of demographic history using the coalescence based approach in MS. One thousand replicate populations were simulated for each of the gene/population combination in Chapter 4, giving a total of 192 sets (4 models x 16 genes x 3 populations). Tajima's *D* was calculated for each replicate population using sample-stats, a program supplied with MS. For each simulation scenario, the mean value of Tajima's *D* (Dsim) and the standard deviation (SDsim) across the 1000 replicates were calculated. The observed values of Tajima's *D* (from Chapter 4) for each gene/population were used to calculate (Dobs-Dsim)/SDsim. If the simulated values of Tajima's *D* closely matched the observed values, then the model that generated those simulated values could be considered to be a reasonable model for the demographic history of the population.

For each population, Africa, Europe and India, the values of (Dobs-Dsim)/SDsim were plotted along a linear scale for each model (Figure 5.2)

a)



b)



(Figure legend overleaf)

c)

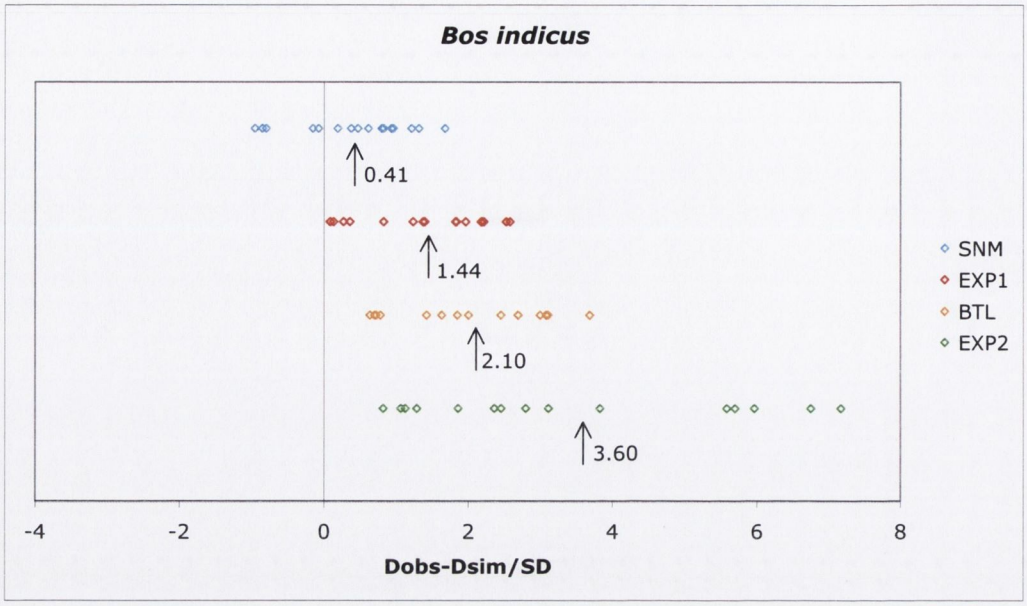


Figure 5.2 Distribution of (Dobs-Dsim)/SDsim for each of the four models of demographic history.

(And preceding page.) a) African *Bos taurus*, b) European *Bos taurus* and c) *Bos indicus*. The models are the standard neutral model, which has a constant population size (blue), the expansion model 1 (EXP1), the bottleneck model (BTL) and the expansion model 2 (EXP2). The mean value of (Dobs-Dsim)/SDsim for each model in each population is indicated by an arrow.

Table 5.1 Summary of (Dobs-Dsim)/SDsim for each of the four models for each population.

Population[1]	Model	Min (Dobs-Dsim)/SDsim	Max (Dobs-Dsim)/SDsim	Range of (Dobs-Dsim)/SDsim	Mean (Dobs-Dsim)/SDsim
AF	SNM	-2.25	1.86	4.11	-0.06
AF	EXP1	-1.29	2.69	3.98	0.76
AF	BTL	-0.87	3.66	4.53	1.25
AF	EXP2	-1.37	6.02	7.39	1.8
EU	SNM	-1.77	1.54	3.31	-0.41
EU	EXP1	-0.73	2.18	2.91	0.66
EU	BTL	-0.57	3.16	3.73	1.31
EU	EXP2	-0.56	6.84	7.4	2.24
IND	SNM	-0.95	1.67	2.62	0.41
IND	EXP1	0.08	2.58	2.5	1.44
IND	BTL	0.64	3.68	3.04	2.1
IND	EXP2	0.83	7.17	6.34	3.6

1. AF, African *Bos taurus*; EU, European *Bos taurus*; IND, Indian *Bos indicus*.

In all three populations, the mean value of (Dobs-Dsim)/SDsim is closest to zero for simulated populations generated using the standard neutral model (African *Bos taurus*: -0.06, European *Bos taurus*: -0.41, *Bos indicus*: 0.41). The same trend is observed in all populations, with expansion model 1 having the second closest mean to zero, the bottleneck model having the third closest mean to zero and the expansion model 2 (EXP2) having the greatest difference between mean (Dobs-Dsim)/SDsim and zero (Table 5.1).

Furthermore, the range of values of (Dobs-Dsim)/SDsim can be considered; a smaller range indicates a more consistent fit to model. EXP2 had the widest range of values among the loci sampled in all three populations. Combined with the observation that it has the greatest difference between the mean (Dobs-Dsim)/SDsim and zero, this suggests that EXP2 is the worst of the three demographic models in terms of describing the variation of observed in these data.

Bos indicus results are consistently different from both of the *Bos taurus* populations, with the mean values always the most positive of the three populations. Only the SNM had values of (Dobs-Dsim)/SDsim below zero. For all four models, the range of values of (Dobs-Dsim)/SDsim was always smallest for the *Bos indicus* population. The Wilcoxon Signed Rank test was used to assess the significance of the differences between populations (SPSS, ver.11 SPSS Inc.) (Table 5.2). No correction was made for multiple testing. TLR5 was excluded from this analysis; TLR5 was not polymorphic in either *Bos taurus* population and so no simulations were carried out to model this gene in these populations.

(Dobs-Dsim)/SDsim was significantly higher in *Bos indicus* than *Bos taurus* in five out of eight taurine-indicine comparisons ($p<0.05$). These include all of the comparisons between European *Bos taurus* and *Bos indicus* and one comparison between African *Bos taurus* and *Bos indicus* (EXP2). Additionally, two more comparisons between African *Bos taurus* and *Bos indicus* (EXP1 and BTL) approached significance ($p<0.2$), further supporting an underlying difference between taurine and indicine demographics. In contrast, none of the comparisons between the two *Bos taurus* and *Bos taurus* populations was significant, or even approached significance. Coalescent-based analysis therefore supports a common demographic history for both taurine populations.

Table 5.2 Wilcoxon Signed Rank test results.

	AF-EU	AF-EU	AF-IND	AF-IND	EU-IND	EU-IND
Model	Z	Significance	Z	Significance	Z	Significance
SNM	-0.852	0.394	-0.795	0.427	-2.215	0.027
EXP1	0.0	1.0	-1.363	0.173	-2.045	0.041
BTL	-0.114	0.91	-1.647	0.10	-2.101	0.036
EXP2	-0.511	0.609	-2.329	0.020	-2.385	0.017

For each model, the (Dobs-Dsim)/SDsim values were compared between populations in a pairwise manner. The test statistic, Z, and the two tailed significance values are shown. Tests that were significant at the $p<0.05$ level are shown in red.

5.3.2 Bayesian analysis (BEAST)

5.3.2.1 Substitution Model

When running BEAST analyses, the nucleotide substitution model must be specified. The HKY model (Hasegawa, Kishino, Yano 85) and the GTR (General Time Reversible) model are available in BEAUTi and these can be further refined by including rate variation between sites and/or invariant sites in the model. Modeltest (Posada and Crandall 1998) was used to assess the most appropriate substitution model to be used in the BEAST analysis using both hLRT (hierarchical likelihood ratio test) and AIC (Akaike Information Content) approaches (Table 5.3). Rate variation is indicated by G, and invariant sites are indicated by I in Table 5.3.

Nearly half (15/32) of the tests performed favour a HKY-based model. Of the remaining tests, five favour TVM, four favour K81, three favour each of F81 and K80, and just a single test favours each of JC and GTR. HKY is the most commonly favoured model by both the hLRT and the AIC, but AIC favours the basic HKY model, while hLRT favours the HKY+I+G. The basic HKY model of substitution was selected for subsequent BEAST analysis as it has been reported that the AIC method out-performs hLRT (Posada and Buckley 2004).

Table 5.3 Modeltest results

Gene	*Bos taurus* (hLRT)	*Bos taurus* (AIC)	*Bos indicus* (hLRT)	*Bos indicus* (AIC)
ART4	HKY	HKY	HKY+I+G	TVM+I
CD2	K80+I	TVMef+I	K80+I	TVMef+I
FEZL	F81	HKY	HKY	HKY
IL2	HKY+I+G	K81uf+I	HKY+I+G	HKY+I
IL13	JC+I	HKY+I	K80+I	K81uf
TNFAIP8L1	HKY	K81uf	F81	HKY
TRIF	HKY+I+G	GTR+I	K81uf	TVM+I
TYROBP	F81+I	HKY+I	HKY+I+G	TVM+I

Results of the Modeltest analyses by the hierarchical likelihood ratio test method and the Akaike Information Content method. The models are as follows: HKY = Hasegawa, Kishino, Yano 85; K81 = Kimura 81(two transversion-parameter model 1); K80 = Kimura 80 (K2P); F81 = Felsenstein 81; JC = Jukes and Cantor; TVM = Tranversional model and GTR = General Time Reversible. I indicates invariant sites and G indicates rate variation among sites.

5.3.2.2 Bayesian Skyline Plots

BEAST offers several advantages over the coalescence-based analyses carried out using MS. First, it is possible to combine multiple sequence regions without introducing errors or wildly violating assumptions. Second, the exploration of demographic parameters is constrained in MS by the models that are input – here only four models of differing demographic shape are investigated as representatives of a vastly larger number of plausible histories. The MCMC method implemented in BEAST has no such constraint and gives a much more complete investigation of parameter space. Third, the simulation analysis relies on goodness of fit to a single test statistic (Tajima's *D*), whereas BEAST estimates effective population size through phylogeny history directly from the coalescent.

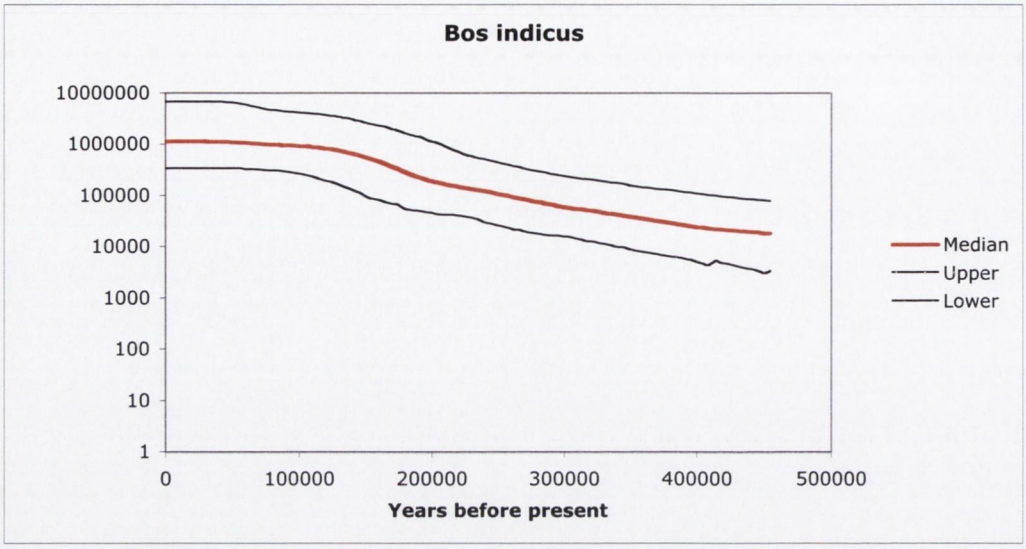
Using BEAST it is possible to combine multiple sequence alignments into a single partitioned file on which the analysis is run. This is particularly useful in this study as the number of polymorphic sites in each alignment was relatively small. By combining eight loci, a more informative analysis could be carried out.

Bayesian Skyline Plots (BSPs) were generated using BEAST for *Bos indicus* (Figure 5.3a)) and *Bos taurus* (Figure 5.3(b)). The *Bos taurus* population was subsequently subdivided into African *Bos taurus* (Figure 5.3(c)) and European *Bos taurus* (Figure 5.3(d)). The BSP is a probabilistic estimate of the effective population size of the population from the present (at the left side of the plot) backwards in time. The units of

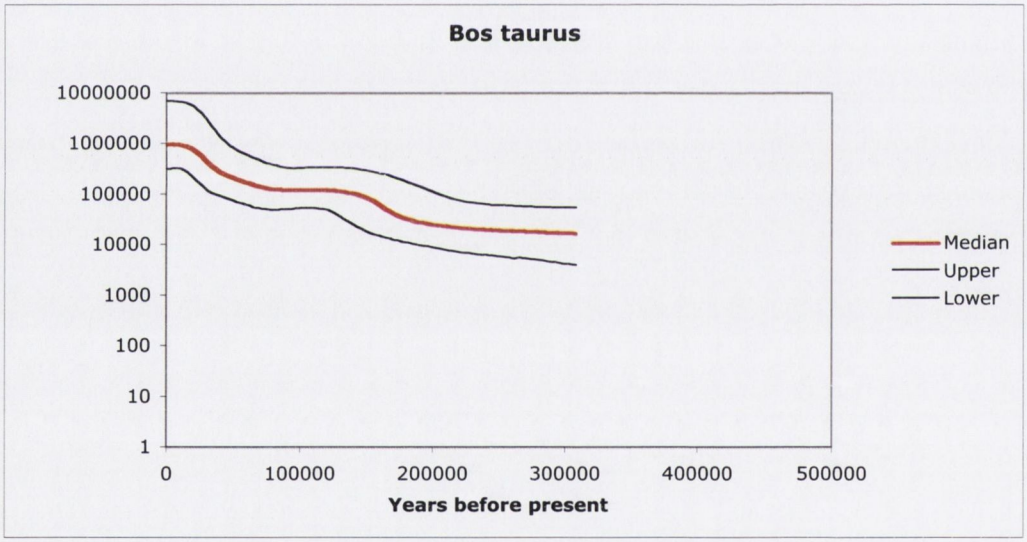
both the x-axis (time) and the y-axis (population size) of a BSP are dependent on the units of the mutation rate specified in BEAUti. The mutation rate used in the analyses was 10^{-8} changes per site per generation (Liu et al. 2006) and assuming a generation time of 5 years (Finlay 2006; Finlay et al. 2007). The time on the x-axis is given in generations, and therefore must be adjusted to years by multiplying by generation time in years. In the case of diploid data such as was used here, the population size on the y-axis is given by $2N_eg$, where N_e is the effective population size and g is the generation time. In this case, the mutation rate used was measured in generations, so g=1. Figure 5.3 shows the corrected x-axis and y-axis values, indicating time in years and N_e, respectively.

All of the BSPs indicate demographic expansion in the history of the populations. The *Bos indicus* BSP is suggestive of an older, more gradual pre-domestication expansion, while the combined *Bos taurus* analysis, as well as the analyses of the subdivided populations are suggestive of an additional more recent, rapid expansion. As the sample size of this group was greater than that of the *Bos indicus* and the African *Bos taurus*, an additional analysis was carried out with sequences from 11 European *Bos taurus* to check whether the larger sample size was influencing the apparently different results between the three samples. Those individuals having haplotypes that clusters with *Bos indicus* haplotypes (see Figure 4.5) were excluded, and additional animals were randomly removed to bring the total to 11. The main features of the BSP were preserved (Figure 5.3(e)).

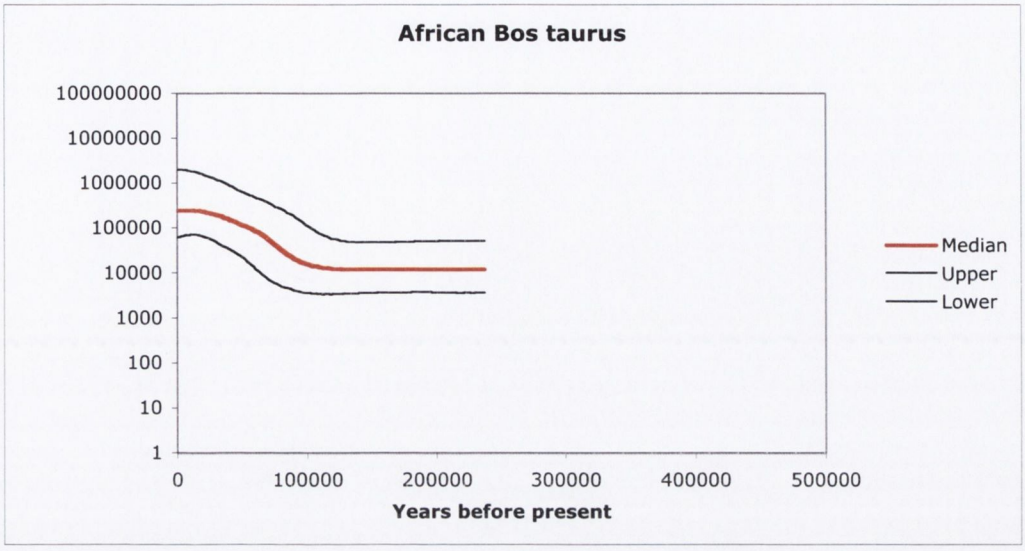
a) *Bos indicus*



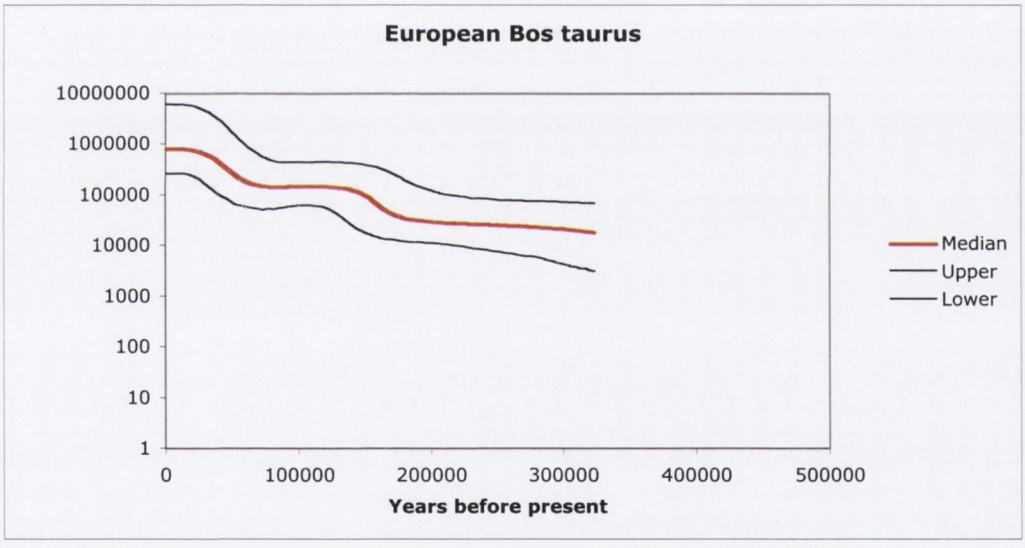
b) *Bos taurus*



c) African *Bos taurus*



d) European *Bos taurus*



e) European *Bos taurus* (n=11)

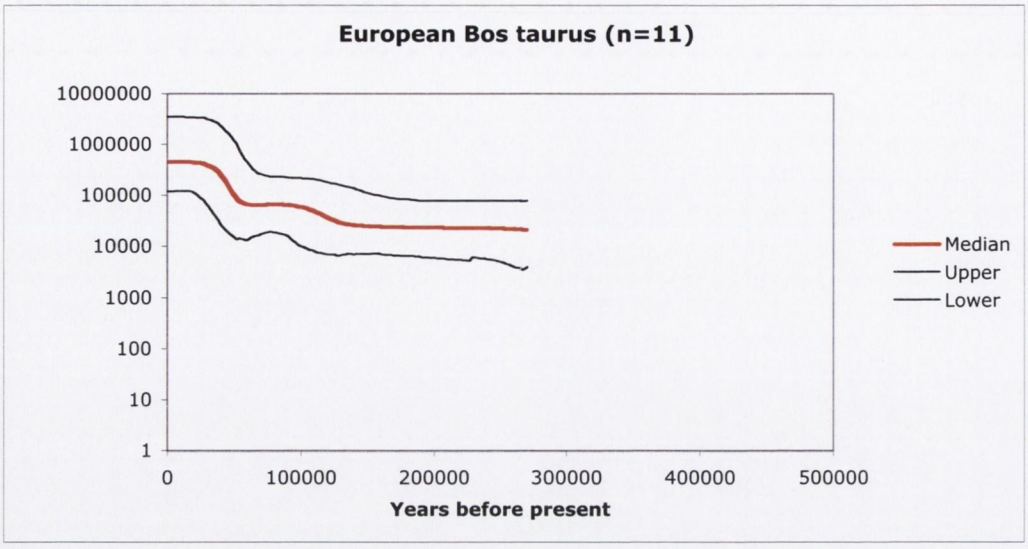


Figure 5.3 Bayesian Skyline Plots for (a) *Bos indicus*; (b) All *Bos taurus*; (c) African *Bos taurus*; (d) European *Bos taurus*; (e) Reduced European *Bos taurus* data (n=11).

(And preceding pages.) The median value of N_e, the effective population size is shown in red. The upper and lower limits of the 95% highest posterior density (HPD) are indicated by black lines.

5.4 Discussion

5.4.1 Different demographic histories for continental populations

The analyses of demographic history described here have been carried out on populations defined at the continental level, that is, an African *Bos taurus* population, a European *Bos taurus* population and a *Bos indicus* population. These three samples were chosen to represent the products of three separate domestication events as previously postulated from molecular data (Bradley et al. 1996; Loftus et al. 1994). Bovine demographic parameter space was explored in two ways: first, using coalescent simulations of loci constrained by four different population histories followed by comparison of observed and simulated summary statistics; and second by using the Bayesian MCMC approach coded in BEAST to estimate probability densities for effective population size through time. Interestingly, several lines of evidence suggest that the demographic history of the *Bos indicus* population is markedly different from either of the *Bos taurus* populations.

The simulated populations that were generated under four different models of demographic history using a coalescent method showed marked differences in their relationships to actual sequence data across the three continental samples. Figure 5.2 shows a distribution of the values of the scaled difference between observed and simulated Tajima's *D* values for each locus under each model and gives a clear visualisation of how well each simulated demographic model fits the observed data. *Bos indicus* values for each locus tend to be skewed to the positive for each model (zero scores would indicate a perfect fit), and this tendency to the positive seems more accentuated than in either *Bos taurus* sample. In fact, distributions of values of (Dobs-Dsim)/SDsim for *Bos indicus* are significantly different to European *Bos taurus* for all four models and also different to African *Bos taurus* for one model (EXP2) (Wilcoxon Signed Rank test; $p<0.05$). At a more basic level, nucleotide diversity is higher within *Bos indicus*; in 11 out of 16 loci, π is greatest in *Bos indicus*, and in eight of 16 loci θ_W is greatest in *Bos indicus*. Haplotype diversity is also consistently higher with 11 of 16 loci giving a highest value in this sample. This strongly suggests divergence in demographic histories between the two taxa. Conversely, no significant difference was found between the distributions of values of (Dobs-Dsim)/SDsim in the African and European *Bos taurus* populations.

The networks of the genes sequences also show that there is more diversity in the *Bos indicus* population than in either the African *Bos taurus*, or the European *Bos taurus*. *Bos*

taurus populations tend to have fewer haplotypes, which are often closely related, possibly indicative of a population expansion. On the other hand, *Bos indicus* networks show a greater number of lower frequency haplotypes, which are very divergent from each other. This is particularly evident for those genes with larger amount of sequence data. (See FEZL, IL2, CD2, TYROBP in Chapter 4). Additionally, where there are outlying *Bos taurus* nodes these tend to occur in the Mediterranean, Near Eastern or African samples; regions where some low levels of zebu introgression are known to have occurred.

Goodness of fit of the Tajima's *D* test statistic for the sequenced loci versus those simulated under different models allows some evaluation of whether features of these models are likely to be accurate. A common supposed domestic history is that it featured a limited sampling from the wild into initial domestic herds; these initial domesticates may then have persisted as small populations for some time in the nascent Neolithic cultures; however a profound population expansion subsequently took place, leading eventually to the hundreds of millions head of livestock which are currently alive. Indeed, a number of earlier lines of genetic evidence seem to support a population bottleneck followed by expansion within cattle and other livestock.

A common feature of domestic ruminant mitochondrial DNA phylogenies is that of discrete star-like clusters of sequences that are suggestive of a history of population expansion (MacHugh and Bradley 2001). In cattle this is particularly evident in networks of mtDNA haplotypes constructed from *Bos taurus*. In Europe and Africa a small number of haplotypes have numerical and topological predominance and are surrounded by numerous derivative singletons and doubletons. These have been argued as representing primary haplotypes captured from the wild, with minor haplotypes representing the results of postdomestic accumulation of mutations (Troy et al. 2001). Zebu mitochondrial DNA networks also show two predominant haplotypes with derivatives (Magee et al. 2007); this pattern is suggestive of population expansion from a narrow base. Lastly, Finlay et al. (2007) examined mtDNA datasets from four domestic *Bovidae* (cattle, mithan, yak and water buffalo) using the Bayesian approach to the coalescent implemented in BEAST and compared these to two wild bovid samples (bison and African buffalo). In each domestic a sharp recent increase in effective population size was estimated within the timeframe of domestic history. Neither wild bovine showed this phenomenon. In the data presented here, there are clusters within several European and African networks, which are reminiscent of the mtDNA signatures interpreted as

limited capture followed by expansion. However, these are markedly absent from *Bos indicus* samples.

With the four demographic histories simulated here, both taurine samples display a distribution of observed Tajima's D values that, on average, fall between those generated by a constant population size of 10^5 (SNM) and a history involving a limited domestication of 1000 animals from a wild pool of 10,000 followed by an expansion to a modern effective population of 100,000 (EXP1). Poor fits are obvious to the other two models: with that of the bottleneck followed by expansion (EXP2) being more divergent to that to a history of bottleneck followed by restoration of a larger population (BTL). Thus this analysis seems to favour the absence of an extended population bottleneck but does not rule out a population expansion model in *Bos taurus,*

Both European and African loci tend to give negative values for Tajima's *D*, which is somewhat indicative of expanding demography. This mirrors results found in human samples, where they are interpreted as the result of significant population expansion (Akey et al. 2004). Some contrast is evident in the examination of Tajima's *D* in *Bos indicus* samples. Here the majority of values are positive (11 out of 16). Also, when these observed values are compared to simulated distributions of the statistic, the observed values are higher than expected, with the exception of a minority of values under the model of constant population size. A comparison of the performance of the different models indicates that the constant population size model (SNM) is the closest fit, with each model incorporating either population expansion or a bottleneck showing a markedly poorer fit. This is surprising given that both of these are presumed features of domestication history.

The Bayesian skyline plots also point to this anomaly. Nevertheless, there is a definite recent sharp population expansion evident in the *Bos taurus* analyses, which is absent in the *Bos indicus* skyline plot. It must be noted that the calibration of demographic behaviour is subject to wide errors. Nevertheless, this expansion could be consistent with one brought by the Near Eastern domestication process. Interestingly, in accordance with the lack of fit with models incorporating a bottleneck above, estimated past population sizes from BEAST are high with median values prior to recent expansion in excess of 100,000. The MS analysis also favours the models with larger ancestral population size over the model with a small pre-domestication population size (EXP2). Notably, the *Bos indicus* sample skyline estimates a consistent, high population size, which is of the order of a million, supporting the Tajima's *D* and MS data.

These analyses raise several interesting issues with respect to domestic history, for example the nature of a population bottleneck associated with domestication. Aurochs were larger than modern cattle and were probably aggressive animals, given that killing an aurochs was a sign of courage in older populations. Therefore it would have seemed reasonable to imagine that only a small number of individuals were domesticated, and that these small numbers gave rise to the gigantic global population of cattle that exist today. However, the data in this work does not strongly support, if at all, a severe bottleneck in the demographic history of cattle. This is particularly true of *Bos indicus*.

These data agree with previous findings from MHC allele diversity by Vila et al. (2005), where they strongly argue that the allele spectrum in many domestic species today are not depleted in diversity as might be expected from a traditional view of early domestic history. Additionally, it has been noted before that levels of protein polymorphism in cattle, as assayed by classical methods, are typical for mammals (Lenstra and Bradley 1999). However, it must be noted that mtDNA diversity in cattle and many domesticates is limited and does resemble a very limited sample of that which would have been present in wild progenitors. A Bayesian skyline of cattle demographic history estimated from mtDNA data estimates a pre-expansion effective population size of the order of 10,000 (Finlay et al. 2007). Perhaps this discrepancy reflects a sexual asymmetry in cattle reproduction.

It is already known that bovine maternal and paternal patterns of diversity can be markedly discordant. For example, cattle herds in Africa show, almost universally, mtDNA sequences from a single *Bos taurus* haplotype family (Troy et al. 2001). However, many genomes are profoundly indicine in both autosomal and Y chromosome markers (Hanotte et al. 2002; Hanotte et al. 2000; MacHugh et al. 1997). Indeed Götherström et al (2005) have suggested that the European domestication process may have involved substantial input from male wild aurochs and not necessarily females. This highlights an important distinction. Incorporation of a mitochondrial DNA chromosome into the domestic gene pool always requires a domestication of an individual animal – an animal that is born wild must be tamed. The size and power of the aurochs is such that this type of event may have been very limited throughout the several thousand year period when domestic cattle herds and wild oxen co-existed in Europe. However, it is feasible that the gene pool resulting from restricted numbers of initially tamed animals was substantially enhanced by repeated crossing of domestic females with wild bulls – it is known, for example that this is normal practice in another domestic bovine, mithan (a

domesticated species whose wild relative is Gaur). This would conceivably give a demographic profile through time that is intermediate between that of models SNM and EXP1 above – as indicated by the actual test statistic values.

In contrast, the statistics for the Indian animals seem to radically exclude either a bottleneck or a classic population expansion. In fact, the Tajima's *D* values tend toward the positive and are in excess of those expected under a constant population size of 10^5. This points toward some complexity in the domestication of *Bos indicus.* Postive Tajima's *D* values (which contrast with the negative tendency in sampled human loci; a signature of past expansion in that species) are often interpreted as indicative of balancing selection. A good number of the loci sampled here are immunologically important genes where selective pressures due to infectious disease challenge could be a factor. However, given that this is a general observation of high values within zebu, balancing selection is an unlikely cause. A more likely scenario is a demographic effect, perhaps the incorporation of two or more divergent strains of South Asian *Bos primigenius* into the domestic pool, which would mimic balancing selection.

Some geographical and genetic complexity within *Bos indicus* domestication does seem likely from prior work. Magee et al. (2007) describe that two clusters of indicine mtDNA haplotypes exist and that these have a non-random distribution when samples from east and west are compared. This is argued as suggesting that more than one domestication event may have given rise to *Bos indicus.* Domestic cattle first appear in the Baluchistan early agricultural centre, Mehrgarh, 8,000-7,000 BP. This is in the context of other domesticates (goats, barley and wheat) which probably diffused eastwards from the Fertile Crescent (Fuller 2006). However, artistic representations and thoracic vertebrae with morphology typical of zebu have allowed Meadow (1987) (reviewed in Fuller 2006) to argue that the early cattle were *Bos indicus.* A transition in the nature of *Bos* bone finds (numbers, size) also points towards this as a plausible centre for domestication of zebu. By the third millennium BC, indicine cattle had spread from their centre of origin and were strong components of important Indus Valley civilisation sites,e.g.Harappa and Mohenjo-daro (Allchin and Allchin 1968; Meadow 1996). These are all part of a northwestern South Asia regional Neotlithic – one of five distinct Neolithic local cecntres proposed by Fuller (Fuller 2006). Overall the evidence for a domestication giving rise to *Bos indicus* here is strong. Additional centres for cattle domestication in the subcontinent are possible. Finds of large *Bos* bones suggest the survival of wild cattle populations into

the Neolithic in South India and possibly the Ganges region further East (reviewed in Fuller 2006).

Direct archaeological evidence for alternative domestication centres for *Bos indicus* is not very strong. The South Indian Neolithic features distinctive ashmounds that are thought to have been produced by the burning of cattle dung (Misra et al. 2001). These have been argued as sites of cattle pens which may have served as enclosures for capture and taming of aurochs 5,000 years ago (Allchin and Allchin 1974). That these may have given rise to a separate strain of zebu is indicated by early representations that seem to depict cattle with longer horns and more delicate limbs in comparison to those represented in early Harappan seals. However, this may simply reflect artistic preferences and allows only weak inference.

Mitochondrial DNA phylogeography gives a clear East-West distinction within zebu haplotype diversity. However it is not possible to judge geographical distributions of the zebu locus variants presented here – the sample is neither large enough nor sufficiently widespread. Fuller (Fuller 2006) has identified five different early agricultural traditions in India, each with distinct putative centres. These each have differing suites of domestic plants and animals and certainly include a number of separate indigenous domestication processes. A further study could profitably examine cattle samples from each of these regions for the haplotypes given here and work toward unravelling the complex orgins of *Bos indicus*.

6 Conclusions

Cattle are important domesticated species that underwent major epidemiological changes at the time of domestication and has certainly faced a range of selective pressures, both artificial and natural. The work presented here has taken complementary approaches to address the question of signatures of selection in the bovine genome. Microsatellite variation and SNP data have been used to identify such signatures of selection. Additionally, cattle demographic history has been considered. This is a vital aspect of any study of selection because demographic changes in a population result in changes genetic variation that may be confused with signals of selection.

The approaches taken here can divided into two types: genome-wide searches for signatures of selection and candidate gene approaches. The former approach does not require any prior knowledge of candidate genes and has the advantage of an inherent empirical distribution from which outliers can be selected. The later approach requires that candidates be selected bases on some predefined criteria. In this study candidate genes were identified from i) PAML analyses (Freeman et al. In Prep.; Lynn et al. 2005). : ART4, CD2, IL2, IL13 and TYROBP; ii) function: TLRs and PRF1 in Chapter 4 and KIR genes in Chapter 3 were chosen based on their role in immune response; and iii) QTL/microsatellite variation: TRIF and TNFAIP8L1 were selected based on their function in immune response and their position in a region of low micrsatellite diversity close to a QTL for trypanotolerence. In this study, the candidate gene approach has been the more fruitful. Two strong candidates (IL2 and IL13) for having been evolving under the influence of positive selection were identified by the neutrality tests in Chapter 4, and additionally other genes showed patterns of diversity that warrant further investigation. In particular, TRIF emerges as a candidate gene for trypanotolerence in African cattle. Analyses of microsatellite varaition on BTA18 lend some support to the theory that the KIR genes are subject to balancing selection. However, microsatellite data is noisy, and further analyses using a resequencing based approach could strengthen this finding.

In theory, genome-wide approaches should yield interesting data but in practice, the quality and quantity of data available are not yet sufficient to produce reliable results in cattle. In the future, studies carried out along the lines of that in Chapter 2 may yield more interesting and dependable results. Large sample populations will be necessary and ideally multiple populations would be available as the work in Chapter 4 indicates that

selection pressures are likely to be specific to populations and not shared across all populations.

A major finding of this work is that the demographic history and population structure of *Bos indicus* is different to that of *Bos taurus*; analyses of both demographic history and genetic diversity data are consistent with there being greater diversity in the *Bos indicus* population. Linkage disequilibrium is typically greater in *Bos indicus* than in either of the *Bos taurus* populations, with the notable exception of TRIF, as discussed in Chapter 4. Haplotype diversity is also greater in the *Bos indicus* population than either *Bos taurus* population for 10 of the 15 genes examined. Networks of haplotypes clearly illustrate that greater diversity is seen in *Bos indicus*, both in terms of the number of haplotypes observed in the sample (Table 4.4) and in the number of mutations separating haplotypes (Figure 4.5). These data could be explained by population structure within *Bos indicus*. It has previously been suggested that the modern *Bos indicus* might be derived from geographically disparate domestication events (Magee et al. 2007) and thus that some substructure may exist within the modern *Bos indicus* population. Indeed, the demographic analyses of Chapter 5 also suggest that the demographic history of *Bos indicus* is markedly different from the two *Bos taurus* populations which showed similar trends in both the goodness of fit tests and the Bayesian Skyline Plots. The skyline plot obtained from the *Bos indicus* sample suggests an older expansion then either of the *Bos taurus* populations. This could be consistent with the *Bos indicus* sample population exhibiting some substructure.

The effects of demographic history on genetic variation must be considered during any tests for signs of selection. Neutrality tests such as those carried out in Chapter 4 may reject the null hypothesis of neutrality in the absence of any selective pressure on the loci being tested if events in the demographic history have resulted in patterns of genetic variation that mimic those expected to be caused by selection. This issue is particularly pertinent when there is reason to think that the species in question has undergone such changes in its history. For domesticated species such as cattle, it is assumed that the species has undergone a bottleneck at the time of domestication and that subsequently the population has expanded. In addition, some variation between modern cattle breeds must exist, though these breed distinctions are relatively new and this substructure is probably too recent to have a major effect on genetic variation at the level of SNPs.

With the exception of the taurine-indicine differences highlighted above, none of the results presented here are suggestive of any major demographic trends in the data: none of the neutrality tests exhibit major departures from expectation and where deviations occur they do so in a pattern that is gene- or population-specific. If a major demographic change was affecting results, one would expect instead to see a consistent, directional deviation from neutrality. The demographic analyses of Chapter 5 are in concordance with this view, favouring the less extreme models of demographic history in each case.

7 References

Aguilar, A., G. Roemer, S. Debenham, M. Binns, D. Garcelon, and R.K. Wayne. 2004. High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proc Natl Acad Sci U S A* **101:** 3490-3494.

Akey, J.M., M.A. Eberle, M.J. Rieder, C.S. Carlson, M.D. Shriver, D.A. Nickerson, and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* **2:** e286.

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 Res* **12:** 1805-1814.

Allchin, B. and F.R. Allchin. 1968. *The birth of Indian civilization: India and Pakistan before 500 BC*. Penguin, Harmondsworth, UK.

Allchin, F.R. and B. Allchin. 1974. Some new thoughts on Indian cattle. In *South Asian archaeology* (eds. J.E. van Lohuizen-de Leeuw and J.N. Ubaghs), pp. 71-77. E. J. Brill, Leiden.

Anisimova, M. and Z. Yang. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol* **24:** 1219-1228.

Antao, T., A. Beja-Pereira, and G. Luikart. 2007. MODELER4SIMCOAL2: a user-friendly, extensible modeler of demography and linked loci for coalescent simulations. *Bioinformatics* **23:** 1848-1850.

Ardell, D.H. 2004. SCANMS: adjusting for multiple comparisons in sliding window neutrality tests. *Bioinformatics* **20:** 1986-1988.

Asthana, S., S. Schmidt, and S. Sunyaev. 2005. A limited role for balancing selection. *Trends Genet* **21:** 30-32.

Balducci, E., L.G. Micossi, E. Soldaini, and R. Rappuoli. 2007. Expression and selective up-regulation of toxin-related mono ADP-ribosyltransferases by pathogen-associated molecular patterns in alveolar epithelial cells. *FEBS Lett* **581:** 4199-4204.

Bancroft, A.J., A.N. McKenzie, and R.K. Grencis. 1998. A critical role for IL-13 in resistance to intestinal nematode infection. *J Immunol* **160:** 3453-3461.

Bauer DuMont, V. and C.F. Aquadro. 2005. Multiple signatures of positive selection downstream of notch on the X chromosome in Drosophila melanogaster. *Genetics* **171:** 639-653.

Beja-Pereira, A., D. Caramelli, C. Lalueza-Fox, C. Vernesi, N. Ferrand, A. Casoli, F. Goyache, L.J. Royo, S. Conti, M. Lari, A. Martini, L. Ouragh, A. Magid, A. Atash, A. Zsolnai, P. Boscato, C. Triantaphylidis, K. Ploumi, L. Sineo, F. Mallegni, P. Taberlet, G. Erhardt, L. Sampietro, J. Bertranpetit, G. Barbujani, G. Luikart, and G. Bertorelle. 2006. The origin of European cattle: evidence from modern and ancient DNA. *Proc Natl Acad Sci U S A* **103:** 8113-8118.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Bökönyi, S. 1974. Development of early stock rearing in the Near East. *Nature* **264:** 19-23.

Borca, M.V., G.F. Kutish, C.L. Afonso, P. Irusta, C. Carrillo, A. Brun, M. Sussman, and D.L. Rock. 1994. An African swine fever virus gene with similarity to the T-lymphocyte surface antigen CD2 mediates hemadsorption. *Virology* **199:** 463-468.

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.

Bradley, D.G. and P. Cunningham. 1999. Genetic aspects of domestication, common breeds and their origins. In *The genetics of cattle* (eds. R. Fries and A. Ruvinsky), pp. 15-31. CAB International, Wallingford, UK.

Bradley, D.G., D.E. MacHugh, P. Cunningham, and R.T. Loftus. 1996. Mitochondrial diversity and the origins of African and European cattle. *Proc Natl Acad Sci U S A* **93:** 5131-5135.

Bradley, D.G. and D.A. Magee. 2006. Genetics and the origins of domestic cattle. In *Documenting domestication. New genetic and archaeological paradigms* (eds. M.A. Zeder D.G. Bradley E. Emshwiller, and B.D. Smith), pp. 317-328. University of California Press, Berkeley and Los Angeles, California.

Broman, K.W. 2001. Estimation of allele frequencies with data on sibships. *Genet Epidemiol* **20:** 307-315.

Brossay, A., F. Hube, T. Moreau, P. Bardos, and H. Watier. 2003. Porcine CD58: cDNA cloning and molecular dissection of the porcine CD58-human CD2 interface. *Biochem Biophys Res Commun* **309:** 992-998.

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

Bustamante, C.D., A. Fledel-Alon, S. Williamson, R. Nielsen, M.T. Hubisz, S. Glanowski, D.M. Tanenbaum, T.J. White, J.J. Sninsky, R.D. Hernandez, D. Civello, M.D. Adams, M. Cargill, and A.G. Clark. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153-1157.

Campbell, K.S. and M. Colonna. 1999. DAP12: a key accessory protein for relaying signals by natural killer cell receptors. *Int J Biochem Cell Biol* **31:** 631-636.

Cargill, E.J. and J.E. Womack. 2007. Detection of polymorphisms in bovine toll-like receptors 3, 7, 8, and 9. *Genomics* **89:** 745-755.

Carlson, C.S., D.J. Thomas, M.A. Eberle, J.E. Swanson, R.J. Livingston, M.J. Rieder, and D.A. Nickerson. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* **15:** 1553-1565.

Casa, A.M., S.E. Mitchell, M.T. Hamblin, H. Sun, J.E. Bowers, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2005. Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. *Theor Appl Genet* **111:** 23-30.

Clark, A.G., S. Glanowski, R. Nielsen, P.D. Thomas, A. Kejariwal, M.A. Todd, D.M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T.J. White, J.J. Sninsky, M.D. Adams, and M. Cargill. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302:** 1960-1963.

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

Cymbron, T., R.T. Loftus, M.I. Malheiro, and D.G. Bradley. 1999. Mitochondrial sequence variation suggests an African influence in Portuguese cattle. *Proc Biol Sci* **266:** 597-603.

Depaulis, F. and M. Veuille. 1998. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol Biol Evol* **15:** 1788-1790.

Di Rienzo, A., A.C. Peterson, J.C. Garza, A.M. Valdes, M. Slatkin, and N.B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci U S A* **91:** 3166-3170.

Drummond, A.J., G.K. Nicholls, A.G. Rodrigo, and W. Solomon. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161:** 1307-1320.

Drummond, A.J. and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7:** 214.

Drummond, A.J., A. Rambaut, B. Shapiro, and O.G. Pybus. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22:** 1185-1192.

Dustin, M.L., D.E. Golan, D.M. Zhu, J.M. Miller, W. Meier, E.A. Davies, and P.A. van der Merwe. 1997. Low affinity interaction of human or rat T cell adhesion molecule CD2 with its ligand aligns adhering membranes to achieve high physiological affinity. *J Biol Chem* **272:** 30889-30898.

Edwards, C.J., R. Bollongino, A. Scheu, A. Chamberlain, A. Tresset, J.D. Vigne, J.F. Baird, G. Larson, S.Y. Ho, T.H. Heupink, B. Shapiro, A.R. Freeman, M.G. Thomas, R.M. Arbogast, B. Arndt, L. Bartosiewicz, N. Benecke, M. Budja, L. Chaix, A.M. Choyke, E. Coqueugniot, H.J. Dohle, H. Goldner, S. Hartz, D. Helmer, B. Herzig, H. Hongo, M. Mashkour, M. Ozdogan, E. Pucher, G. Roth, S. Schade-Lindig, U. Schmolcke, R.J. Schulting, E. Stephan, H.P. Uerpmann, I. Voros, B. Voytek, D.G. Bradley, and J. Burger. 2007. Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proc Biol Sci* **274:** 1377-1385.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5:** 435-445.

Epstein, H. 1971. *The origin of the domestic cattle of Africa.* Africana, New York.

Epstein, H. and I.L. Mason. 1984. Cattle. In *Evolution of domesticated animals* (ed. I.L. Mason), pp. 6-27. Longman, London.

Ewens, W.J. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* **3:** 87-112.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

Fay, J.C. and C.I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155:** 1405-1413.

Felius, M. 1995. *Cattle breeds: An encyclopedia.* Trafalgar Square.

Felsenstein, J. 1989. PHYLIP: phylogeny inference package. *Cladistics* **5:** 164-166.

Finlay, E.K. 2006. Bovini Mitochondrial DNA : Demographic History and Recombination. PhD thesis. University of Dublin.

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

Freeman, A.R. 2003. Admixture analysis of domesticated cattle. PhD thesis. University of Dublin.

Freeman, A.R., D.J. Lynn, C. Murray, and D.G. Bradley. In Prep. Detecting the effects of selection at the population level in bovine immune genes.

Fu, Y.X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147:** 915-925.

Fu, Y.X. and W.H. Li. 1993. Statistical tests of neutrality of mutations. *Genetics* **133:** 693-709.

Fuller, D.Q. 2006. Agricultural origins and frontiers in South Asia: A working synthesis. *Journal of World Prehistory* **20:** 1-86.

Gaffen, S.L. and K.D. Liu. 2004. Overview of interleukin-2 function, production and clinical applications. *Cytokine* **28:** 109-123.

Gibbs, B.F. 2005. Human basophils as effectors and immunomodulators of allergic inflammation and innate immunity. *Clin Exp Med* **5:** 43-49.

Gilad, Y., C.D. Bustamante, D. Lancet, and S. Paabo. 2003. Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am J Hum Genet* **73:** 489-501.

Gilad, Y., D. Segre, K. Skorecki, M.W. Nachman, D. Lancet, and D. Sharon. 2000. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nat Genet* **26:** 221-224.

Goldammer, T., H. Zerbe, A. Molenaar, H.J. Schuberth, R.M. Brunner, S.R. Kata, and H.M. Seyfert. 2004. Mastitis increases mammary mRNA abundance of beta-defensin 5, toll-like-receptor 2 (TLR2), and TLR4 but not TLR9 in cattle. *Clin Diagn Lab Immunol* **11:** 174-185.

Goldstein, D.B., A. Ruiz Linares, L.L. Cavalli-Sforza, and M.W. Feldman. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A* **92:** 6723-6727.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8:** 195-202.

Götherström, A., C. Anderung, L. Hellborg, R. Elburg, C. Smith, D.G. Bradley, and H. Ellegren. 2005. Cattle domestication in the Near East was followed by hybridization with aurochs bulls in Europe. *Proc Biol Sci* **272:** 2345-2350.

Gu, Z., X. Zhao, N. Li, and C. Wu. 2007. Complete sequence of the yak (Bos grunniens) mitochondrial genome and its evolutionary relationship with other ruminants. *Mol Phylogenet Evol* **42:** 248-255.

Gubin, A.N., J.M. Njoroge, U. Wojda, S.D. Pack, M. Rios, M.E. Reid, and J.L. Miller. 2000. Identification of the dombrock blood group glycoprotein as a polymorphic member of the ADP-ribosyltransferase gene family. *Blood* **96:** 2621-2627.

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

Hanotte, O., C.L. Tawah, D.G. Bradley, M. Okomo, Y. Verjee, J. Ochieng, and J.E. Rege. 2000. Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-saharan African cattle breeds. *Mol Ecol* **9:** 387-396.

Hedrick, P.W. and G. Thomson. 1983. Evidence for balancing selection at HLA. *Genetics* **104:** 449-456.

Hubbard, T.J., B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S.C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. 2007. Ensembl 2007. *Nucleic Acids Res* **35:** D610-617.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. In *Oxford surveys in Evolutionary Biology* (eds. D. Futuyma and J. Antonovics), pp. 1-43. Oxford University Press, Oxford.

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

Ihara, N., A. Takasuga, K. Mizoshita, H. Takeda, M. Sugimoto, Y. Mizoguchi, T. Hirano, T. Itoh, T. Watanabe, K.M. Reed, W.M. Snelling, S.M. Kappes, C.W. Beattie, G.L. Bennett, and Y. Sugimoto. 2004. A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Res* **14:** 1987-1998.

Iraqi, F., K. Sekikawa, J. Rowlands, and A. Teale. 2001. Susceptibility of tumour necrosis factor-alpha genetically deficient mice to Trypanosoma congolense infection. *Parasite Immunol* **23:** 445-451.

Jiggins, F.M. and K.W. Kim. 2007. A screen for immunity genes evolving under positive selection in Drosophila. *J Evol Biol* **20:** 965-970.

Kato, K., M. Koyanagi, H. Okada, T. Takanashi, Y.W. Wong, A.F. Williams, K. Okumura, and H. Yagita. 1992. CD48 is a counter-receptor for mouse CD2 and is involved in T cell activation. *J Exp Med* **176:** 1241-1249.

Kayser, M., S. Brauer, and M. Stoneking. 2003. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol Biol Evol* **20:** 893-900.

Kelley, J., L. Walter, and J. Trowsdale. 2005. Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet* **1:** 129-139.

Kelley, J.L., J. Madeoy, J.C. Calhoun, W. Swanson, and J.M. Akey. 2006. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* **16:** 980-989.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217:** 624-626.

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

Kingman, J.F.C. 1982. On the genealogy of large populations. *Journal of Applied Probability* **19A:** 27-43.

Kreitman, M. and A. Di Rienzo. 2004. Balancing claims for balancing selection. *Trends Genet* **20:** 300-304.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5:** 150-163.

Lanier, L.L., B.C. Corliss, J. Wu, C. Leong, and J.H. Phillips. 1998. Immunoreceptor DAP12 bearing a tyrosine-based activation motif is involved in activating NK cells. *Nature* **391:** 703-707.

Lee, B.T., T.W. Tan, and S. Ranganathan. 2003. MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences. *Nucleic Acids Res* **31:** 3533-3536.

Lenstra, J.A. and D.G. Bradley. 1999. Systematics and phylogeny of cattle. In *The genetics of cattle* (eds. R. Fries and A. Ruvinsky), pp. 1-14. CAB International, Wallingford, UK.

Liu, G.E., L.K. Matukumalli, T.S. Sonstegard, L.L. Shade, and C.P. Van Tassell. 2006. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. *BMC Genomics* **7:** 140.

Loftus, R.T., D.E. MacHugh, D.G. Bradley, P.M. Sharp, and P. Cunningham. 1994. Evidence for two independent domestications of cattle. *Proc Natl Acad Sci U S A* **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. *Nat Rev Genet* **4:** 981-994.

Luo, M., X. Mao, and F.A. Plummer. 2005. Identification of four novel HLA-B alleles, B*1590, B*1591, B*2726, and B*4705, from an East African population by high-resolution sequence-based typing. *Tissue Antigens* **65:** 187-191.

Lynn, D.J., A.R. Freeman, C. Murray, and D.G. Bradley. 2005. A genomics approach to the detection of positive selection in cattle: adaptive evolution of the T-cell and natural killer cell-surface protein CD2. *Genetics* **170:** 1189-1196.

MacHugh, D.E. 1996. Molecular biogeography and genetic structure of domesticated cattle. PhD thesis. University of Dublin.

MacHugh, D.E. and D.G. Bradley. 2001. Livestock genetic origins: goats buck the trend. *Proc Natl Acad Sci U S A* **98:** 5382-5384.

MacHugh, D.E., R.T. Loftus, D.G. Bradley, P.M. Sharp, and P. Cunningham. 1994. Microsatellite DNA variation within and among European cattle breeds. *Proc Biol Sci* **256:** 25-31.

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.

Magee, D.A., H. Mannen, and D.G. Bradley. 2007. Duality in *Bos indicus* mitochondrial DNA diversity: Support for geographical complexity in zebu domestication. In *The evolution and history of human populations in South Asia. Inter-disciplinary studies in archaeology, biological anthropology, linguistics and genetics.* (eds. M.D. Petraglia and B. Allchin), pp. 385-391. Springer.

Martin, A.M., E.M. Freitas, C.S. Witt, and F.T. Christiansen. 2000. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* **51:** 268-280.

Martin, M.P., G. Nelson, J.H. Lee, F. Pellett, X. Gao, J. Wade, M.J. Wilson, J. Trowsdale, D. Gladman, and M. Carrington. 2002. Cutting edge: susceptibility to psoriatic arthritis: influence of activating killer Ig-like receptor genes in the absence of specific HLA-C alleles. *J Immunol* **169:** 2818-2822.

Matsuoka, Y., S.E. Mitchell, S. Kresovich, M. Goodman, and J. Doebley. 2002. Microsatellites in Zea - variability, patterns of mutations, and use for evolutionary studies. *Theor Appl Genet* **104:** 436-450.

Matthews, D.E., R.A. Hessler, and T.W. O'Brien. 1978. Rapid evolutionary divergence of proteins in mammalian mitochondrial ribosomes. *FEBS Lett* **86:** 76-80.

Mattiangeli, V., A.W. Ryan, R. McManus, and D.G. Bradley. 2006. A genome-wide approach to identify genetic loci with a signature of natural selection in the Irish population. *Genome Biol* **7:** R74.

McKenzie, G.J., A. Bancroft, R.K. Grencis, and A.N. McKenzie. 1998. A distinct role for interleukin-13 in Th2-cell-mediated immune responses. *Curr Biol* **8:** 339-342.

McPeek, M.S., X. Wu, and C. Ober. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* **60:** 359-367.

McQueen, K.L., B.T. Wilhelm, K.D. Harden, and D.L. Mager. 2002. Evolution of NK receptors: a single Ly49 and multiple KIR genes in the cow. *Eur J Immunol* **32:** 810-817.

Meadow, R. 1987. Faunal exploitation patterns in eastern Iran and Baluchistan: A review of recent investigations. In *Oreintalia Iosephi Tucci Memoriae Dicata* (eds. G. Gnoli and L. Lanciotti), pp. 881-916. Instituto Italiano per il Medio ed Estermo Oreinte, Rome.

Meadow, R. 1996. The origins and spread of agriculture and pastoralism in northwestern South Asia. In *The origins and spread of agriculture and pastoralism in Eurasia.* (ed. D.R. Harris), pp. 390-412. UCL Press, London.

Meadow, R.H. 1993. Animal domestication in the Middle East: A revised view from the Eastern margin. In *Harappan civilization* (ed. G. Possehl), pp. 295-320. Oxford and IBH, New Dehli.

Mears, J.A., M.R. Sharma, R.R. Gutell, A.S. McCook, P.E. Richardson, T.R. Caulfield, R.K. Agrawal, and S.C. Harvey. 2006. A structural model for the large subunit of the mammalian mitochondrial ribosome. *J Mol Biol* **358:** 193-212.

Minch, E. MICROSAT http://hpgl.stanford.edu/projects/microsat/microsat.html

Minty, A., P. Chalon, J.M. Derocq, X. Dumont, J.C. Guillemot, M. Kaghad, C. Labit, P. Leplatois, P. Liauzun, B. Miloux, and et al. 1993. Interleukin-13 is a new human

lymphokine regulating inflammatory and immune responses. *Nature* **362:** 248-250.

Misra, V.D., J.N. Pal, and M.C. Gupta. 2001. Excavation at Tokwa: A Neolithic-Chalcolithic settlement. *Pragdhara* **11:** 59-72.

Morton, H.C., R.J. Pleass, A.K. Storset, E. Dissen, J.L. Williams, P. Brandtzaeg, and J.M. Woof. 2004. Cloning and characterization of an immunoglobulin A Fc receptor from cattle. *Immunology* **111:** 204-211.

Nair, S., J.T. Williams, A. Brockman, L. Paiphun, M. Mayxay, P.N. Newton, J.P. Guthmann, F.M. Smithuis, T.T. Hien, N.J. White, F. Nosten, and T.J. Anderson. 2003. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. *Mol Biol Evol* **20:** 1526-1536.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nickerson, D.A., V.O. Tobe, and S.L. Taylor. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* **25:** 2745-2751.

Nielsen, R., C. Bustamante, A.G. Clark, S. Glanowski, T.B. Sackton, M.J. Hubisz, A. Fledel-Alon, D.M. Tanenbaum, D. Civello, T.J. White, J.S. J, M.D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3:** e170.

Nielsen, R., I. Hellmann, M. Hubisz, C. Bustamante, and A.G. Clark. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* **8:** 857-868.

Niimura, Y. and M. Nei. 2007. Extensive gains and losses of olfactory receptor genes in Mammalian evolution. *PLoS ONE* **2:** e708.

Noe-Nygaard, N., T.D. Price, and S.U. Hede. 2005. Diet of aurochs and early cattle in southern Scandinavia: evidence from 15N and 13C stable isotopes. *Journal of Archaeological Science* **32:** 855-871.

Norman, P.J., M.A. Cook, B.S. Carey, C.V. Carrington, D.H. Verity, K. Hameed, D.D. Ramdath, D. Chandanayingyong, M. Leppert, H.A. Stephens, and R.W. Vaughan. 2004. SNP haplotypes and allele frequencies show evidence for disruptive and balancing selection in the human leukocyte receptor complex. *Immunogenetics* **56:** 225-237.

O'Brien, T.W. 2003. Properties of human mitochondrial ribosomes. *IUBMB Life* **55:** 505-513.

O'Neill, L.A. and A.G. Bowie. 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat Rev Immunol* **7:** 353-364.

Ohta, T. and M. Kimura. 1973. The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. *Genetic Research* **22:** 201-204.

Parham, P. 2005. MHC class I molecules and KIRs in human history, health and survival. *Nat Rev Immunol* **5:** 201-214.

Park, S.D.E. 2002. Trypanotolerance in West African cattle and the population genetic effects of selection. PhD thesis. University of Dublin.

Paterson, S. 1998. Evidence for balancing selection at the major histocompatibility complex in a free-living ruminant. *J Hered* **89:** 289-294.

Payseur, B.A. and A.D. Cutter. 2006. Integrating patterns of polymorphism at SNPs and STRs. *Trends Genet* **22:** 424-429.

Payseur, B.A., A.D. Cutter, and M.W. Nachman. 2002. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol Biol Evol* **19:** 1143-1153.

Pearce-Duvet, J.M. 2006. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. *Biol Rev Camb Philos Soc* **81:** 369-382.

Pietromonaco, S.F., R.A. Hessler, and T.W. O'Brien. 1986. Evolution of proteins in mammalian cytoplasmic and mitochondrial ribosomes. *J Mol Evol* **24:** 110-117.

Pipkin, M.E. and J. Lieberman. 2007. Delivering the kiss of death: progress on understanding how perforin works. *Curr Opin Immunol* **19:** 301-308.

Posada, D. and T.R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol* **53:** 793-808.

Posada, D. and K.A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14:** 817-818.

Pybus, O.G., A. Rambaut, and P.H. Harvey. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* **155:** 1429-1437.

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

Roach, J.C., G. Glusman, L. Rowen, A. Kaur, M.K. Purcell, K.D. Smith, L.E. Hood, and A. Aderem. 2005. The evolution of vertebrate Toll-like receptors. *Proc Natl Acad Sci U S A* **102:** 9577-9582.

Rodriguez, J.M., R.J. Yanez, F. Almazan, E. Vinuela, and J.F. Rodriguez. 1993. African swine fever virus encodes a CD2 homolog responsible for the adhesion of erythrocytes to infected cells. *J Virol* **67:** 5312-5320.

Rozas, J., J.C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19:** 2496-2497.

Rozen, S. and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365-386.

Schlenke, T.A. and D.J. Begun. 2003. Natural selection drives Drosophila immune system evolution. *Genetics* **164:** 1471-1480.

Schlenke, T.A. and D.J. Begun. 2005. Linkage disequilibrium and recent selection at three immunity receptor loci in Drosophila simulans. *Genetics* **169:** 2013-2022.

Schlotterer, C., M. Kauer, and D. Dieringer. 2004. Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality. *Proc Biol Sci* **271:** 869-874.

Schroder, K., P.J. Hertzog, T. Ravasi, and D.A. Hume. 2004. Interferon-gamma: an overview of signals, mechanisms and functions. *J Leukoc Biol* **75:** 163-189.

Selvaraj, P., M.L. Plunkett, M. Dustin, M.E. Sanders, S. Shaw, and T.A. Springer. 1987. The T lymphocyte glycoprotein CD2 binds the cell surface ligand LFA-3. *Nature* **326:** 400-403.

Shapiro, B., A.J. Drummond, A. Rambaut, M.C. Wilson, P.E. Matheus, A.V. Sher, O.G. Pybus, M.T. Gilbert, I. Barnes, J. Binladen, E. Willerslev, A.J. Hansen, G.F. Baryshnikov, J.A. Burns, S. Davydov, J.C. Driver, D.G. Froese, C.R. Harington, G. Keddie, P. Kosintsev, M.L. Kunz, L.D. Martin, R.O. Stephenson, J. Storer, R. Tedford, S. Zimov, and A. Cooper. 2004. Rise and fall of the Beringian steppe bison. *Science* **306:** 1561-1565.

Shimojima, M., Y. Nishimura, T. Miyazawa, K. Kato, K. Nakamura, Y. Izumiya, H. Akashi, and Y. Tohya. 2002. A feline CD2 homologue interacts with human red blood cells. *Immunology* **105:** 360-366.

Single, R.M., M.P. Martin, X. Gao, D. Meyer, M. Yeager, J.R. Kidd, K.K. Kidd, and M. Carrington. 2007. Global diversity and evidence for coevolution of KIR and HLA. *Nat Genet* **39:** 1114-1119.

Smirnova, I., M.T. Hamblin, C. McBride, B. Beutler, and A. Di Rienzo. 2001. Excess of rare amino acid polymorphisms in the Toll-like receptor 4 in humans. *Genetics* **158:** 1657-1664.

Stephens, M. and P. Scheet. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76:** 449-462.

Stephens, M., J.S. Sloan, P.D. Robertson, P. Scheet, and D.A. Nickerson. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat Genet* **38:** 375-381.

Stepp, S.E., R. Dufourcq-Lagelouse, F. Le Deist, S. Bhawan, S. Certain, P.A. Mathew, J.I. Henter, M. Bennett, A. Fischer, G. de Saint Basile, and V. Kumar. 1999. Perforin gene defects in familial hemophagocytic lymphohistiocytosis. *Science* **286:** 1957-1959.

Storset, A.K., I.O. Slettedal, J.L. Williams, A. Law, and E. Dissen. 2003. Natural killer cell receptors in cattle: a bovine killer cell immunoglobulin-like receptor multigene family contains members with divergent signaling motifs. *Eur J Immunol* **33:** 980-990.

Strimmer, K. and O.G. Pybus. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol* **18:** 2298-2305.

Sugimoto, M., A. Fujikawa, J.E. Womack, and Y. Sugimoto. 2006. Evidence that bovine forebrain embryonic zinc finger-like gene influences immune response associated with mastitis resistance. *Proc Natl Acad Sci U S A* **103:** 6454-6459.

Swofford, D.L. 2002 Paup*: Phylogenetic Analysis Using Parsimony (and other methods) http://www.sinauer.com/detail.php?id=8060

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-595.

Takeda, K., T. Kaisho, and S. Akira. 2003. Toll-like receptors. *Annu Rev Immunol* **21:** 335-376.

Tangye, S.G., J.H. Phillips, and L.L. Lanier. 2000. The CD2-subset of the Ig superfamily of cell surface molecules: receptor-ligand pairs expressed by NK cells and other immune cells. *Semin Immunol* **12:** 149-157.

Teshima, K.M., G. Coop, and M. Przeworski. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16:** 702-712.

Trowsdale, J., R. Barten, A. Haude, C.A. Stewart, S. Beck, and M.J. Wilson. 2001. The genomic context of natural killer receptor extended gene families. *Immunol Rev* **181:** 20-38.

Troy, C.S., D.E. MacHugh, J.F. Bailey, D.A. Magee, R.T. Loftus, P. Cunningham, A.T. Chamberlain, B.C. Sykes, and D.G. Bradley. 2001. Genetic evidence for Near-Eastern origins of European cattle. *Nature* **410:** 1088-1091.

Uhrberg, M. 2005. The KIR gene family: life in the fast lane of evolution. *Eur J Immunol* **35:** 10-15.

van Baren, M.J. and P. Heutink. 2004. The PCR suite. *Bioinformatics* **20:** 591-593.

Vila, C., J. Seddon, and H. Ellegren. 2005. Genes of domestic mammals augmented by backcrossing with wild ancestors. *Trends Genet* **21:** 214-218.

Vilches, C. and P. Parham. 2002. KIR: diverse, rapidly evolving receptors of innate and adaptive immunity. *Annu Rev Immunol* **20:** 217-251.

Vincendeau, P. and B. Bouteille. 2006. Immunology and immunopathology of African trypanosomiasis. *An Acad Bras Cienc* **78:** 645-665.

Voight, B.F., S. Kudaravalli, X. Wen, and J.K. Pritchard. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72.

Volz, A., H. Wende, K. Laun, and A. Ziegler. 2001. Genesis of the ILT/LIR/MIR clusters within the human leukocyte receptor complex. *Immunol Rev* **181:** 39-51.

Wang, J. 2002. Protein recognition by cell surface receptors: physiological receptors versus virus interactions. *Trends Biochem Sci* **27:** 122-126.

Watterson, G.A. and R. Anderson. 1978. A Further Analysis of Multiallelic Data. *Genetics* **90:** 207-210.

Wendorf, F. and R. Schild. 1994. Are the early holocene cattle in the eastern Sahara domestic or wild? *Evolutionary Anthropology* **3:** 118-128.

Werling, D. and T.W. Jungi. 2003. TOLL-like receptors linking innate and adaptive immune response. *Vet Immunol Immunopathol* **91:** 1-12.

White, S.N., K.H. Taylor, C.A. Abbey, C.A. Gill, and J.E. Womack. 2003. Haplotype variation in bovine Toll-like receptor 4 and computational prediction of a positively selected ligand-binding domain. *Proc Natl Acad Sci U S A* **100:** 10364-10369.

Appendix A

Polymorphisms observed in cattle sequences for a) ART4, b) CD2, c) FEZL, d) IL2, e) IL13, f) MRPL30, g) MRPS05, h) MRPS14, i)PRF1, j)TLR2, k) TLR4, l) TLR5, m)TLR9 n) TNFAIP8L1, o) TRIF and p) TYROBP. No polymorphic sites were observed in TLR7. The assembly position refers to the position of the polymorphic site in the Phred/phrap assembly. This number has no biological meaning but has been retained for reference to the assembly. This number can also be used to refer to schematic diagrams in figure 4.x The SNP column indicates the nucleotides observed at each position in the sequencing panel. The genomic position is given relative to the A of the start codon (A is position 1). The consensus sequence refers to the consensus sequence from the Phred/phrap assembly. For those polymorphisms that lie within the protein coding region of the gene, the position of the codon in the protein is indicated, whether or not the polymorphism is non-synonymous, and the amino acid(s) that the codon containing the polymorphic side encodes.

a) ART4

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
4490	G/T	-174	T			
4540	C/T	-124	T			
4601	A/G	-63	G			
4620	C/G	-44	C			
4646	C/G	-18	C			
4677	A/G	13	G	5	Yes	R/G
4766	A/G	102	G	34	No	A
4786	G/T	122	G	41	Yes	R/M
4817	C/T	153	T	51	No	S
4845	A/G	181	A			
4903	A/G	239	G			
4947	A/G	283	A			
4956	C/T	292	C			
4957	A/G	293	G			
5005	C/T	341	C			
5011	A/G	347	G			
5039	A/G	375	A			
5066	A/C	402	C			
5158	C/T	494	C			
5201	C/T	537	C			
8422	A/G	3758	G			
8541	A/G	3877	A			
8651	C/T	3987	C			
8701	C/T	4037	T			
8960	G/T	4296	G	126	Yes	G/W
8972	A/G	4308	A	130	Yes	I/V
9221	G/T	4557	G	213	Yes	A/S
9286	A/G	4622	G	234	No	G
9298	A/G	4634	G	238	No	L
9454	A/G	4790	G	290	No	T
9469	A/G	4805	G	295	No	L
9488	A/G	4824	G			
9492	C/G	4828	C			
9495	A/T	4831	T			
9537	C/T	4873	C			
9547	A/G	4883	G			
9569	A/G	4905	G			
9593	A/G	4929	A			

b) CD2

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
2482	C/T	656	C			
2641	C/T	815	C			
2662	A/G	836	G			
2790	C/T	964	T			
7133	C/G	5307	G	151	Yes	Q/E
7149	C/T	5323	T	156	Yes	T/M
7221	A/T	5395	A	180	Yes	K/M
7226	A/G	5400	G	182	Yes	M/V
7248	A/G	5422	G	189	Yes	K/R
7249	C/G	5423	C	189	Yes	N/K
7260	C/T	5434	T	193	Yes	T/M
7261	A/G	5435	A	193	No	T
7271	A/G	5445	G	197	Yes	M/V
11510	C/G	9684	G			
11593	A/C	9767	C			
11668	A/G	9842	G	216	Yes	T/A
11680	A/C	9854	C	220	Yes	I/L
11770	C/T	9944	C			
11876	C/T	10050	T			
15186	A/G	13360	A			
15274	C/G	13448	C			
15455	A/G	13629	A	251	No	A
15480	A/C	13654	A	260	Yes	T/P
15500	A/G	13674	A	266	No	S
15623	A/G	13797	G	307	No	P
15686	A/G	13860	A	328	No	T
15704	C/T	13878	C	334	No	A
15756	A/G	13930	G			
15980	A/G	14154	A			
16337	A/G	14511	A			
16353	A/G	14527	A			
16379	A/G	14553	A			
16393	C/G	14567	C			

c) FEZL

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
476	G/T	-150	T			
516	A/G	-110	G			
674	C/G	48	G	16	No	P
1171	C/T	545	C	182	Yes	P/S
1173	A/G	547	G	182	No	P
1784	C/T	1158	C			
1985	A/G	1359	G			
1993	C/G	1367	C			
2096	C/T	1470	C			
2423	A/G	1797	G			
2453	A/G	1827	G			
2681	A/G	2055	A			
2743	C/T	2117	C			
2820	A/G	2194	A			
2829	A/G	2203	G			
2937	A/G	2311	G			
3909	C/T	3283	C			

d)IL2

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
1760	G/T	-821	G			
2381	A/G	-200	A			
2541	C/T	-40	T			
5283	C/T	2702	T			
5710	C/T	3129	T	111	No	I
6031	A/T	3450	T			
6058	A/G	3477	A			
6107	A/G	3526	A			
6150	C/T	3569	T			
6708	A/T	4127	A			
6750	A/C	4169	A			
6775	C/T	4194	C			
7156	C/T	4575	C			
7170	A/G	4589	A			

e) IL13

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
422	C/G	-578	C			
690	A/C	-310	C			
800	C/T	-200	C			
1019	A/G	19	G	7	Yes	T/A
1176	A/G	176	G			
1204	A/G	204	A			
1285	C/T	285	T			
1370	C/G	370	G			
1448	A/T	448	T			
1493	G/T	493	T			
1500	C/G	500	G			
1619	C/T	619	C			
1657	C/T	657	C			
1813	C/G	813	C			
2161	C/T	1161	C			
2276	C/G	1276	C			
2301	C/T	1301	C			
2342	A/G	1342	G			
2376	C/T	1376	C			
2681	A/T	1681	T			
2966	C/T	1966	C			
3054	A/G	2054	A			
3175	C/T	2175	C			
3233	C/G	2233	C			
3299	A/C	2299	A			
3633	A/G	2633	A			

f) MRPL30

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
2164	C/T	-5791	T			
2318	A/G	-5637	G			
2360	A/G	-5595	A			
15153	C/T	7198	C			
15166	C/T	7211	T			
15183	C/T	7228	C			
15301	A/C	7346	C			
15383	A/T	7428	A	117	Yes	I/L
15397	C/T	7442	C			
15523	A/G	7568	G	137	No	T
15606	A/G	7651	G			

g) MRPS05

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
957	A/C	169	C			
1044	C/G	256	C	34	Yes	T/R
1060	A/G	272	G	39	No	A
3528	G/T	2740	T			
3563	C/G	2775	G			
3595	C/G	2807	G	54	Yes	P/A
3632	C/T	2844	C	66	Yes	A/V
3675	A/G	2887	G	80	No	T
3736	A/C	2948	C			
6814	A/T	6026	A	107	No	G
6853	A/G	6065	A	120	No	R
6855	A/G	6067	A	121	Yes	K/R
6892	C/T	6104	C	133	No	G
6917	A/G	6129	A			
6999	A/G	6211	G			
10016	A/T	9228	A			
10130	C/T	9342	C			
10139	C/G	9351	C			
10171	A/G	9383	G			
10188	A/C	9400	A			
10416	A/G	9628	A			
10508	G/T	9720	T			
10552	A/G	9764	A			
10586	A/G	9798	G			
10607	A/G	9819	G			
10620	C/T	9832	T			
10622	A/G	9834	G			
10672	A/G	9884	G			
10725	A/G	9937	A			
10748	C/T	9960	C			
10777	A/G	9989	G			
10782	C/T	9994	T			

h) MRPS14

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
952	A/G	353	G			
8200	C/T	7601	T			

i) PRF1

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
4329	A/G	1727	G	317	Yes	S/N
4564	A/G	1962	G	395	Yes	M/I
4611	A/G	2009	A	411	Yes	R/K
4836	C/T	2234	T	486	Yes	F/S
4903	A/G	2301	A	508	No	S
5252	G/T	2650	T			

j) TLR2

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
3019	A/G	455	G	152	Yes	R/Q
3166	A/G	602	G	201	Yes	S/N
3195	A/G	631	G	211	Yes	I/V
3245	A/T	681	T	227	Yes	F/L
3509	G/T	945	G	315	No	R
3542	A/T	978	T	326	Yes	H/R
3574	A/G	1010	G	337	Yes	R/K
3593	A/C	1029	A	343	No	I
3778	C/T	1214	C	405	Yes	T/M
3814	A/G	1250	A	417	Yes	N/S
3872	A/C	1308	A	436	No	G
4018	C/T	1454	C	485	Yes	S/F
4068	G/T	1504	T	502	Yes	S/A

k) TLR4

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
9337	A/C	8337	C	347	Yes	A/E
9416	A/G	8416	A	373	No	P
9439	A/G	8439	A	381	Yes	K/R
9450	C/T	8450	T	385	No	L
9464	G/T	8464	T	389	No	G
9818	A/G	8818	A	507	No	Q
9953	C/T	8953	T	552	No	S
10064	C/T	9064	T	589	No	S
10124	C/T	9124	T	609	No	C
10163	C/T	9163	T	622	No	S
10172	C/T	9172	C	625	No	N
10244	A/G	9244	G	649	No	G
10245	A/G	9245	G	650	Yes	V/I
10289	A/C	9289	C	664	No	G
10318	C/T	9318	C	674	Yes	T/I
10325	C/T	9325	T	676	No	D

l) TLR5

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
2179	C/T	2079	C	693	No	Y
2308	C/G	2208	G	736	No	P

m) TLR9

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
4526	A/G	2361	A	787	No	T
4580	C/T	2415	T	805	No	R
4589	A/G	2424	G	808	No	L
4625	A/G	2460	G	820	No	S

n) TNFAIP8L1

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
340	A/G	-760	G			
901	A/G	-199	A			
1554	A/G	454	G	152	Yes	N/D
1670	A/G	570	G			
1679	C/T	579	C			
1805	C/G	705	C			
1883	C/T	783	C			
2183	A/T	1083	A			
2341	C/T	1241	C			
2342	A/T	1242	A			
2586	C/T	1486	C			
2632	A/G	1532	G			

o) TRIF

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
112	C/G	-711	G			
163	C/T	-660	C			
853	C/G	30	C	10	No	G
1068	A/G	245	G	82	Yes	N/S
1303	C/G	480	C	160	No	S
1330	A/G	507	G	169	No	S
1354	C/T	531	C	177	No	S
1626	C/T	803	C	268	Yes	P/L
1676	A/G	853	A	285	Yes	T/A
1687	C/G	864	C	288	Yes	H/Q
1704	C/T	881	T	294	Yes	P/L
1714	C/T	891	C	297	No	S
1785	A/C	962	C	321	Yes	N/T
1849	C/T	1026	T	342	No	A
1851	A/C	1028	A	343	Yes	Q/P
1859	A/C	1036	C	346	Yes	T/P
2105	A/G	1282	G	414	Yes	I/V
2278	C/T	1455	C	471	No	P
2630	A/C	1807	C	589	Yes	K/Q
2662	C/T	1839	C	599	No	H
2769	A/G	1946	G	635	Yes	E/G
2787	A/G	1964	G	641	Yes	E/G
2800	C/T	1977	C	645	No	S
2819	C/T	1996	T	652	Yes	P/S
2837	G/T	2014	G	658	Yes	G/C
2887	A/G	2064	G	674	No	P
2896	A/G	2073	G	677	No	P
2907	A/C	2084	A	681	Yes	Q/P
2929	A/G	2106	G	688	No	P
2950	A/G	2127	G	695	No	P
2984	A/C	2161	A	707	Yes	T/P
2988	C/T	2165	C	708	Yes	P/L
3112	A/G	2289	G	749	No	Q
3121	A/G	2298	G	752	No	E
3130	A/G	2307	G	755	No	T
3324	C/T	2501	C			
3383	A/G	2560	A			
3581	A/G	2758	G			
3583	C/T	2760	T			
3745	A/G	2922	A			
3752	C/T	2929	T			
3836	A/G	3013	A			
3846	A/G	3023	A			
3889	A/G	3066	G			
3895	C/T	3072	C			
3961	G/T	3138	G			
4007	A/G	3184	A			
4016	C/T	3193	C			
4017	A/G	3194	A			
4058	C/T	3235	C			

p) TYROBP

Assembly Position	SNP	Genomic Position	Consensus Seq	Protein Position	Non-Synonymous	Amino acids
180	A/G	-875	G			
258	G/T	-797	T			
338	A/T	-717	T			
717	C/T	-338	C			
767	A/G	-288	A			
795	A/G	-260	G			
905	A/G	-150	G			
1090	C/G	35	C	12	Yes	S/C
1262	A/G	207	G			
1331	A/G	276	G			
1454	A/C	399	A			
1506	A/T	451	T	24	Yes	Q/L
1579	G/T	524	T			
1639	C/T	584	C			
1878	A/G	823	G			
1893	A/C	838	A			
1910	C/G	855	C			
2034	C/G	979	C			
2041	C/T	986	C			
2755	A/C	1700	C			
2875	C/T	1820	C			
3188	A/G	2133	A			
3218	A/G	2163	G			
3434	C/T	2379	C			
3449	A/T	2394	T			
3653	C/T	2598	C	92	No	G
3976	C/G	2921	G			