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.

Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection

Stephen David Edward Park

A thesis submitted to the University of Dublin
for the degree of Doctor of Philosophy

Department of Genetics,
Trinity College,
University of Dublin

June 2002

DECLARATION

This thesis has not been submitted as an exercise for a degree at any other university. Except where stated, the work described therein was carried out by me alone.

I give permission for the Library to lend or copy this thesis upon request.

Signed:

Stephen D. E. Park

June 2002

CONTENTS

FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

TABLES

Chapter 2

Chapter 3

Chapter 4

ACKNOWLEDGEMENTS

I would like first of all to thank my parents for giving their constant support to make this thesis possible, even though they must have wondered if I would ever stop being a student.

I am deeply indebted to Dan Bradley whose remarkable patience, insight and sound advice have seen me through to the end. I couldn't have wished for a better supervisor. To Dave and Ronan, the dual undomesticates thanks for blazing the trail and setting me on the right track. Thanks to Chris, sometime anarchist and expert on bulls who, with Ciaran 'Chairman Gene', taught me to dive and opened up the world of the deep. To Emmy, always on hand for 'just the one' pint, and to Ashie, champion of fish in a cow's world To Jill and John, so long and thanks for all the curry.

My thanks also to thank Niamh Millar, who gave up a summer in New York to run gels in the Bovine Lab, and to Valerie Corr, whose MSc project contributed some of the BTA5 data. To the next generation, Ruth, Dave, Ceiridwen, Brian, who keep the flame alive – best of luck with everything. And to everyone else who has made the Bovine lab such a great place to work. And whatever happens, we will always have La Cave.

There are many other members of the Genetics department for whose help I am very grateful. Thanks are of course due to Dave, Paul and Louis, who were always there when things ran out or went wrong, and who always kept the lab going. I would like to thank Andrew Lloyd and Wolfe cubs Aoife and Karsten for fielding all my questions about the mysteries of bioinformatics and the way of UNIX. I am also very grateful to David, Kasper and Libby from the Bacillus lab for showing me the joys of cloning.

I was very privileged to spend several months at ILRI, thanks to Olivier Hanotte, who looked after me so well. I am especially grateful to Philomeen Nilsson for going out of her way to help at all times, and for making sure I didn't leave without seeing the country. I would also like to thank Simon Kang'a and Joel Mwakaya for all their help with the work, and indeed everyone in Lab 7 for making my stay so enjoyable.

I would also like to acknowledge the assistance of our other collaborators: Dr. Racine Sow of ISRA, Dakar, who organised sample collection in Senegal, and Dr. Balabadi Dao of Sokodé who organised sampling in Togo; Dr. Tom Goldammer, who provided a chromosome library.

Thanks finally to all at No. 29: Jonathan, Adrian, Helen and Colin.

The work described in this thesis was funded by The European Commission (DGXII: INCO Contract No. ERBIC18CT950005).

ABBREVIATIONS

BLAST®	Basic Local Alignment Search Tool
BovR5	5,000 rad bovine radiation hybrid mapping panel
BovR12	12,000 rad bovine radiation hybrid mapping panel
bp	Base pairs (nucleotides)
BP	Years Before present
BTA1, BTA2 *etc.*	Bovine chromosome 1, 2 etc. (abbreviation of *Bos taurus*)
cDNA	Complementary DNA, transcribed from messenger RNA
cM	Centi-Morgans
cR, cR_{5000}, cR_{12000}, *etc.*	Centi-rad radiation hybrid map distances (subscripts show radiation doses used to create mapping panels)
d.f.	Degrees of freedom
$(\delta\mu)^2$	Delta mu squared genetic distance
D_S	Nei's standard genetic distance
EST	Expressed sequence tag
F_{exp}	Predicted homozygosity under selective neutrality
F_{obs}	Sample homozygosity calculated under assumption of HWE
H-F	Holstein-Friesian
HSA1, HSA5 *etc.*	Human chromosome 1, 5 etc. (abbreviation of *Homo sapiens*)
HWE	Hardy-Weinberg equilibrium
IAM	Infinite alleles model of neutral mutation
ILRI	International Livestock Research Institute (Nairobi, Kenya)
kb	Kilobase pairs (nucleotides)
Mb	Megabase pairs (nucleotides)
MDE	Mutation-drift equilibrium
MHC	Major histocompatibility complex
mtDNA	Mitochondrial DNA
NCBI	National Center for Biotechnology Information (USA)
N_e	Effective population size
PCR	Polymerase chain reaction
PCV	Packed [red blood] cell volume
RH	Radiation hybrid
QTL	Quantitative trait locus / loci
SINE	Short interspersed element (repetitive DNA element)
SMM	Stepwise mutation model
SNP	Single nucleotide polymorphism
STS	Sequence tagged site

SUMMARY

A panel of 69 microsatellite loci was used to investigate the population genetic effects of selection at genes on linked, neutral markers. Loci were typed in African and European *Bos taurus* cattle and *Bos indicus* populations from the Indian subcontinent. Where data was already available, this was added to the dataset. The microsatellites included five loci within 3cM of the bovine class IIa MHC region on chromosome 23, eight loci each within a few kilobases of known genes, and 56 anonymous loci not known to be linked to genes. A range of genetic analyses were performed on the data.

Microsatellites linked to the MHC class IIa region show significantly higher gene diversity and numbers of alleles per locus than anonymous loci. Allele frequency distributions for MHC-linked loci are significantly less skewed than those seen for anonymous loci. Genetic distance between distantly related populations is significantly lower when calculated using MHC-linked loci than when using anonymous loci. In contrast, gene-linked loci show significantly lower gene diversity and numbers of alleles per locus than anonymous loci, and significantly more skewed allele distributions. These observations are attributed to the effects of selection. Overdominant selection (heterozygote advantage) is known to act at MHC class II genes, and is predicted to cause all of the effects described. Selection at genes is more likely to favour one particular allele, or to act to purge deleterious mutations. Both models of selection are predicted to have consequences for linked loci that agree with the observations.

Similar population genetic analyses are applied to data from microsatellites from three candidate chromosomal regions identified from a genome scan as potentially responsible for genetic tolerance to trypanosomiasis in West African cattle. It is hypothesised that selection for disease tolerance in cattle from West Africa may have been intense. However, no population genetic evidence is found to suggest that the microsatellite loci investigated have been influenced by selection at nearby genes responsible for trypanotolerance. Possibly the microsatellites are too distant from trypanotolerance genes. The considerable inherent variation between microsatellite loci may also mask the effects of selection.

Strategies are developed to identify new polymorphic loci within candidate trypanotolerance chromosomal regions for potential use in locating trypanotolerance genes. Conservation of gene location and structure between human and cattle is exploited to amplify introns from six previously unmapped bovine genes. Radiation hybrid mapping is

used to confirm that these genes map to the expected region on chromosome 7, and to add additional microsatellite loci to the map. Comparison of sequences of the amplified introns reveals considerable sequence polymorphism. Five of the six introns contain SNPs, and overall SNP frequency is 1 per 261bp. Most SNPs discriminate between zebu and taurine cattle, reflecting the ancient separation of the sub-species.

The potential for identifying SNPs in bovine sequence data in the public domain is also assessed. Six apparent bovine SNP loci are identified in sequences with human homologues mapping to the region of human chromosome corresponding to the BTA7 trypanotolerance QTL region.

CHAPTER 1

GENERAL INTRODUCTION

1.1 CATTLE

1.1.1 Cattle taxonomy

Cattle (*Bos* sp.) belong to the *Bovini* tribe within the *Bovidae* family. There are two recognised sub-species of domesticated cattle: the taurine cattle found in Europe, West Africa and northern Asia, (*B. taurus*) and the zebu cattle found in the Indian sub-continent and northern and eastern Africa (*B. indicus*). The types of cattle show considerable morphological and physiological differences. The most striking difference is that zebu cattle have a hump, either in a cervico-thoracic (6th cervical to the 5th thoracic vertebrae) position or in a thoracic (1st to 9th thoracic vertebrae) position. Zebu cattle also have a large dewlap and a navel flap which allow for heat loss. This, along with a lower basal metabolic rate, lower water requirements, and larger and more active sweat glands reflects the zebu's adaptation to arid climates. In contrast, taurine cattle are adapted to more temperate zones with higher levels of rainfall (Epstein 1971).

1.1.2 Origins of cattle

The wild ancestor of domesticated cattle was the *aurochs*, *Bos primigenius*, which ranged across Eurasia and North Africa. Local speciation within this very broad geographic range gave rise to at least three distinct races of aurochsen: *Bos primigenius primigenius* in Europe and northern Asia, *Bos primigenius opisthonomous* in North Africa, and *Bos primigenius namadicus* in Southern Asia (Felius 1995).

The earliest archaeological evidence of cattle domestication comes from the Neolithic site of Çatal Hüyük in southern central Turkey. Cattle remains from strata dating from approximately 8,400 BP show apparent evidence of domestic manipulation. Remains from strata dating from approximately 7800 BP show a characteristic reduction in size

associated with the domestication process (Perkins 1969). Archaeological evidence suggests that independent domestication of cattle, also during the Neolithic period, occurred in the Mehrgarh region in Baluchistan, southern Pakistan (Meadow 1984; Meadow 1993). The earliest strata from Mehrgarh date from 9000 BP, with cattle becoming the predominant ruminant species by 7000 BP. Osteological analysis indicates that the cattle remains are *Bos indicus* (Meadow 1984; Meadow 1993). This suggests that zebu and taurine cattle were domesticated at sites separated by considerable distance on either side of the Iranian plateau.

Using craniometric and osteological measurements of modern taurine and zebu cattle and aurochsen fossils, Grigson finds that zebu cattle more closely resemble the southern Asiatic aurochs *Bos primigenius namadicus*, while taurine cattle more closely resemble the northern Eurasian aurochs *Bos primigenius primigenius* (Grigson 1978; Grigson 1980). This points to *B. p. namadicus* being the ancestor of zebu cattle and *B. p. primigenius* being the ancestor of taurine cattle.

Strong evidence for different ancestry of taurine and zebu cattle is provided by genetic analysis. A molecular clock approach has been used to estimate divergence dates for the different types of cattle. This method involves comparing differences between the types of cattle with differences between cattle and an outgroup species bison (*Bison* sp.), from which cattle diverged approximately 1.0 million years ago (Hartl *et al.* 1988). Applying this approach to mitochondrial DNA (mtDNA) data, an estimated divergence time of 117,000-275,000 years before present is obtained for taurine and zebu cattle (Bradley *et al.* 1996; Loftus *et al.* 1994). The estimate obtained using microsatellite data suggests a divergence time of 610,000-850,000 years before present (MacHugh *et al.* 1997). These

dates clearly precede domestication during the Neolithic period, and support the view that cattle were domesticated from different sub-species of aurochsen in Europe and Asia.

1.1.3 Expansion of cattle out of the Near East

Domesticated cattle are believed to have been introduced into Europe by advancing pastoralists migrating from the Near East (Epstein and Mason 1984; Troy *et al.* 2001). Cattle appear also to have been brought into Iberia from North Africa via the straits of Gibraltar as evinced by the presence of African mtDNA haplotypes in modern Portuguese cattle breeds (Cymbron *et al.* 1999). Initial introductions were of longhorn cattle. Subsequently, smaller shorthorn cattle emerged, either through local adaptation or as a result further introductions from Mesopotamia, as suggested by Epstein and Mason (1984). **Figure 1.1** shows presumed migration routes (assuming a separate near-eastern origin for shorthorn cattle rather than local adaptation from longhorn animals).

Some taurine cattle originating in the Near East may have been brought to the Indian sub-continent. Artefacts from the Harappan civilisation, which flourished from about 5,000 BP in the Indus Valley in Pakistan, show representations of both zebu and taurine cattle (Kulke and Rothermund 1990). Genetic evidence also suggests a taurine genetic component in modern cattle breeds from North India and Pakistan (MacHugh *et al.* 1997), possibly as a result of ancient introductions of taurine animals. However, the animals have remained morphologically indicine. Zebu cattle are thought to have spread rapidly throughout the Indian sub-continent from their centre of origin in modern Pakistan (**Figure 1.1**), and over the past 4,000 to 5,000 years the only cattle in the sub-continent cattle have been zebu.

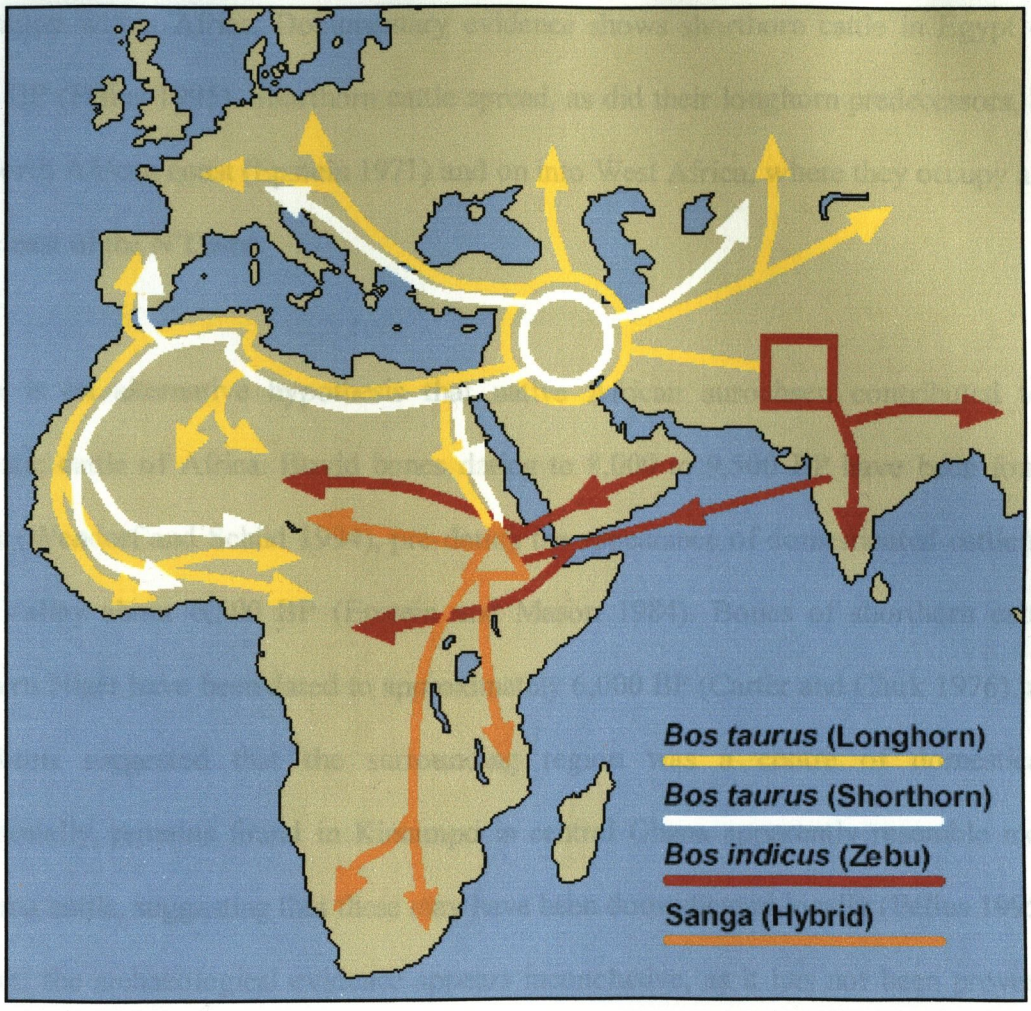

Figure 1.1: Postulated centres of origin and migration patterns of domesticated cattle
Closed shapes represent the centres of origin of the various types of cattle. Arrows represent migration routes. Figure taken from MacHugh (1996) and based originally on Payne (1970) and Epstein and Mason (1984).

1.1.4 Origins of African cattle

The traditional view has been that domesticated taurine cattle were introduced into Africa via Egypt from the Near East around the time that they were introduced into Europe (Epstein 1971; Epstein and Mason 1984). Postulated migration routes are shown in **Figure 1.1**. Once introduced into Egypt, migrating peoples from the Nile valley brought their cattle south through modern Sudan, as well as westwards along the North African coastal region, and from there into West Africa. These longhorn animals gave rise to the N'Dama and derived breeds in West Africa. Subsequently, shorthorn cattle also emerged, possibly through new introductions from the near east in the view of Epstein, or through local

adaptation within Africa. Documentary evidence shows shorthorn cattle in Egypt 4,000 years BP (Felius 1995). Shorthorn cattle spread, as did their longhorn predecessors, along the North African coast (Epstein 1971) and on into West Africa, where they occupy a zone to the east of the N'Dama.

There is an alternative hypothesis that native African aurochsen contributed to the domestic cattle of Africa. Bovid bones dating to 8,000 to 9,500 BP have been found in Egypt (Wendorf and Schild 1994), pre-dating the appearance of domesticated cattle in the Nile Valley about 6,500 BP (Epstein and Mason 1984). Bones of shorthorn cattle in northern Niger have been dated to approximately 6,000 BP (Carter and Clark 1976), and it has been suggested that the surrounding region was a centre of domestication. Additionally, remains found in Kintampo in central Ghana apparently resemble modern N'Dama cattle, suggesting that these may have been domesticated locally (Felius 1995). At present, the archaeological evidence appears inconclusive, as it has not been proved that any of the remains are of domestic animals. Much more compelling is phylogenetic evidence which supports the thesis of a separate African centre of domestication. Calculations based on both nuclear microsatellite data (MacHugh *et al.* 1997) and on mtDNA data (Bradley *et al.* 1996; Troy *et al.* 2001) provide evidence that the separation between taurine cattle originating in the Near East and those of Africa pre-dates the development of agriculture. Calculations based on nuclear microsatellite data suggest a very ancient split between European and African taurine cattle of between 150,000 and 250,000 years BP (MacHugh *et al.* 1997). Estimates from mtDNA are much lower, but again pre-date the rise of agriculture. Using mtDNA the separation has been estimated at 22-26,000 years BP (Bradley *et al.* 1996) and in a separate study at 10,100-37,600 years BP (Troy *et al.* 2001). Strikingly the latter study finds that the African mtDNA haplotypes cluster around a central haplotype that is absent in Europe, and very rare in the Near East

(Troy *et al.* 2001). Collectively, this archaeological and genetic evidence points to domestication of a local wild Aurochs species that was relatively closer to the ancestor of European taurine cattle than that of zebu cattle.

1.1.5 Introduction of zebu cattle into Africa

After the emergence of taurine cattle in Africa, zebu cattle were introduced into Africa from Arabia and the Indian sub-continent (**Figure 1.1**). From about 4,000 BP, cervico-thoracic humped (neck-humped) zebu cattle were introduced into East Africa (Payne 1970), where they were crossed with local taurine cattle to form *sanga* breeds which latterly spread into Southern and Central Africa. Much later, thoracic-humped ('wither-humped') cattle were introduced into East Africa from Arabia following the Arab conquests of 669 A.D. These cattle were dispersed westwards through the Sahel region to the south of the Sahara and north of the forest belt. Hybridisation with local taurine cattle has occurred in this region (**Figure 1.2**) so that most Sahel cattle populations now have a zebu genetic component (MacHugh *et al.* 1997). The extent of this component varies according to the climate and associated trypanosomiasis incidence. Zebu cattle are drought-adapted, whereas taurine cattle are less so. Conversely, the local taurine N'Dama and shorthorn cattle have evolved tolerance to trypanosomiasis, present in humid areas where the tsetse fly vector is abundant (**Figure 1.3**), while zebu cattle are susceptible to the disease. Consequently, populations further north (e.g. in Mauritania) have a substantial zebu genetic component, whereas populations to the south (in e.g. Guinea) have little or none (MacHugh *et al.* 1997).

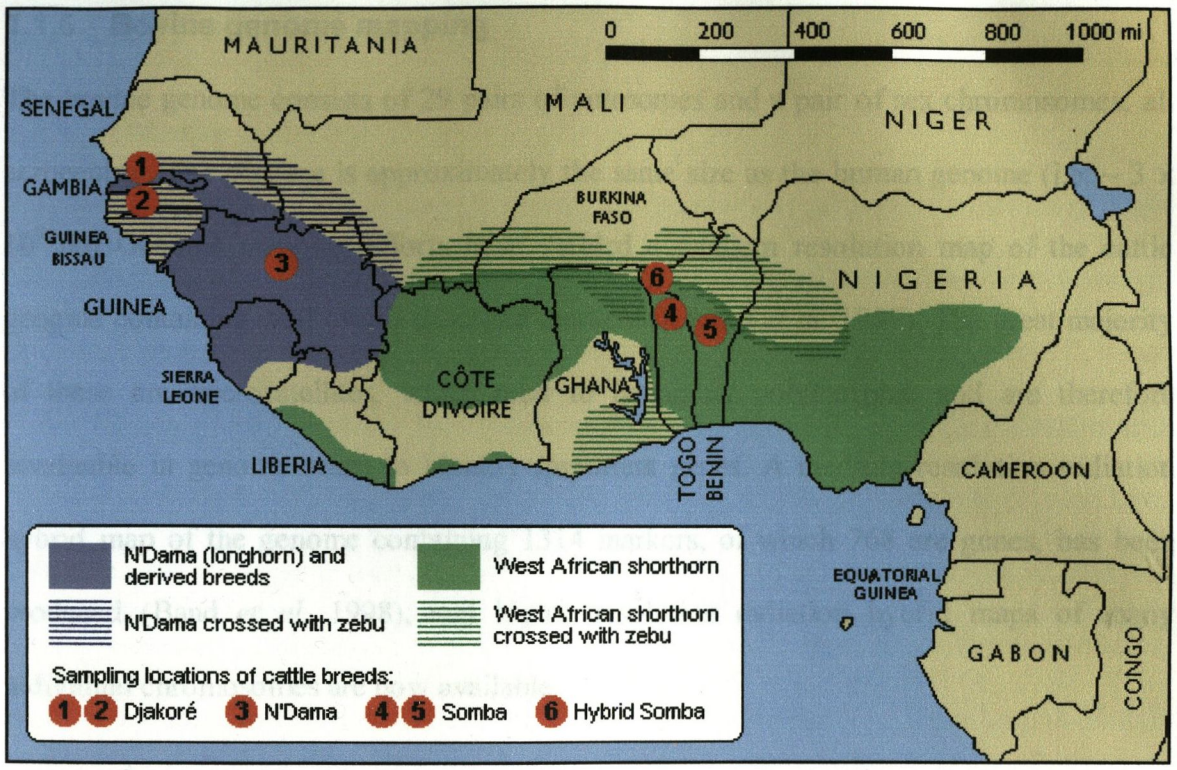

Figure 1.2: Regions of origin and sampling locations of African cattle breeds studied

Adapted from Felius (1995)

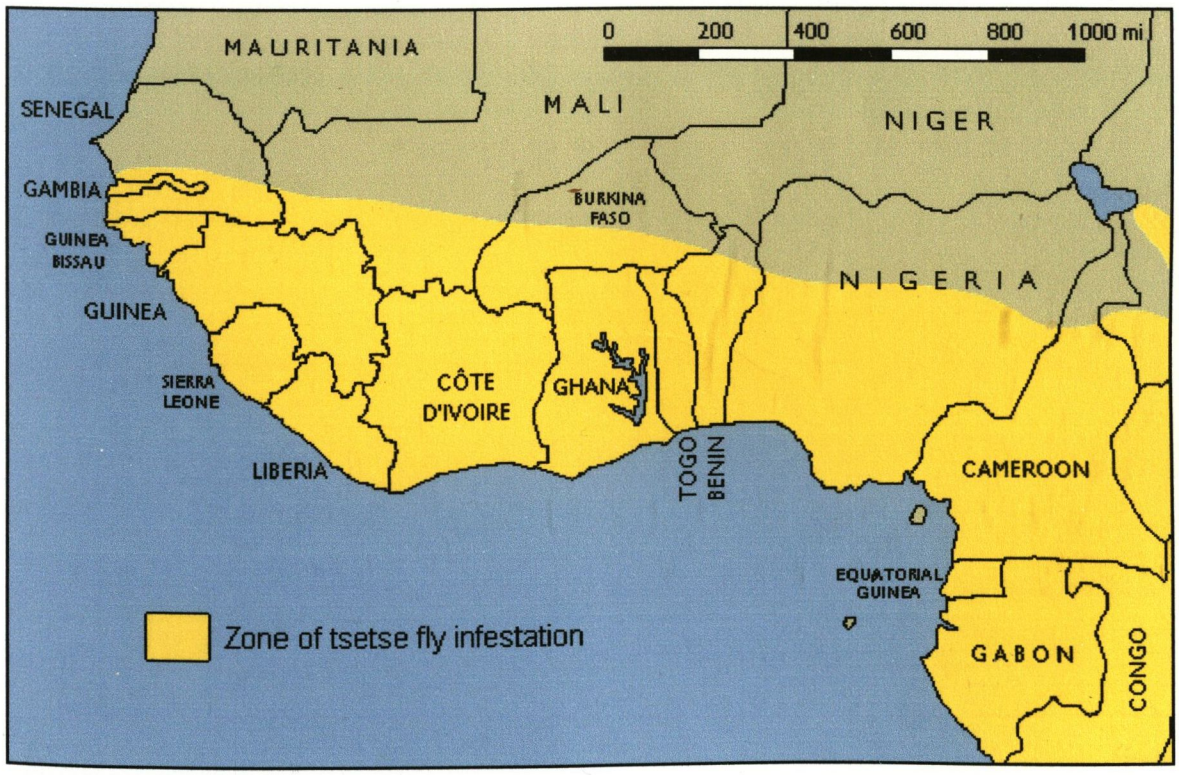

Figure 1.3: Zone of tsetse-borne trypanosomiasis in West Africa

Adapted from Trail *et al.* (1979)

1.1.6 Bovine genome mapping

The bovine genome consists of 29 pairs of autosomes and a pair of sex chromosomes, all acrocentric. The genome is approximately the same size as the human genome (i.e. ~ 3 x 10^9 bp) Genome mapping efforts have yielded a medium resolution map of the cattle genome containing well over a thousand markers (Kappes *et al.* 1997). The great majority of these are microsatellites, which tend to be highly polymorphic and are therefore invaluable in genome scans to identify important genes. A medium resolution radiation hybrid map of the genome containing 1314 markers, of which 768 are genes, has been produced (Band *et al.* 1998), and higher resolution radiation hybrid maps of many individual chromosomes are now available.

1.2 TRYPANOSOMIASIS

1.2.1 Disease overview

Trypanosomiasis is a wasting disease caused by extracellular protozoan parasites (**Figure 1.4**). The disease affects many mammal species, including cattle. The principal trypanosome species pathogenic to cattle in Africa are *Trypanosoma congolense*, *T. vivax*, *T. brucei brucei* and *T. simiae*. All of these are transmitted by the tsetse fly (*Glossina* sp.). Disease symptoms include intermittent fever, anaemia, diarrhoea, and loss of condition. Infection causes failure to thrive and loss of production, and is often fatal (Maré 1988). A recent review by d'Ieteren et al (1998) reports that the disease has a major economic impact in Africa, as it is a major constraint on animal production in humid areas. Humid areas cover 19% of Africa, but contain only 6% of the continent's livestock. It has been estimated that the eradication of tsetse-borne trypanosomiasis, which currently affects approximately 7 million square kilometres, would allow an increase in the cattle population of 33 million.

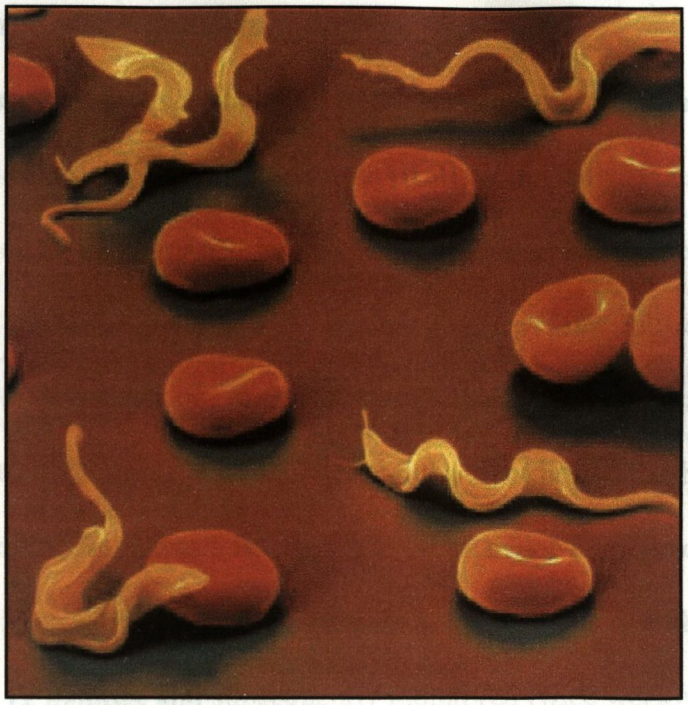

Figure 1.4: *Trypanosoma brucei* parasites

T. brucei parasites shown as yellow; red blood cells shown as red / orange

Figure from Nature, 8th June 2000

1.2.2 Trypanosomiasis control

There are various options for controlling trypanosomiasis including eradication of the tsetse fly, the use of trypanocidal drugs, the development of a vaccine, and the use of cattle which have evolved tolerance to the disease (d'Ieteren *et al.* 1998). Many attempts have been made over decades to eradicate tsetse flies such as habitat destruction, widespread insecticide spraying, the release of sterile flies and the use of traps. These approaches have however proved of limited effect, and may be environmentally damaging. Growing drug resistance among trypanosome populations means that drugs are becoming less effective, whilst their expense also limits usage. Currently prospects of a vaccine are still remote, due to the ability of trypanosomes to change their antigenic variation. The remaining prospect, namely the exploitation of cattle that are naturally tolerant to the disease, is therefore receiving serious attention.

1.2.3 Trypanotolerance gene mapping

A whole genome scan for genes involved in trypanotolerance in cattle has been undertaken at the International Livestock Research Institute (ILRI). The pedigree used in the scan was created by crossing four trypanotolerant N'Dama sires from the Gambia to four trypanosusceptible Boran dams from Kenya to produce four full-sib F1 (first filial) generation families (Hanotte *et al.*, in preparation). From these families, 13 males and 17 females were selected to produce an F2 generation, avoiding full-sib matings. The F2 generation consisted of 203 animals belonging to seven major families of 22 to 39 animals and a number of minor families. The F2 animals were challenged with trypanosomes (*T. congolense*) at age 12 months, and subsequently monitored twice weekly over 150 days for body weight, parasite load and packed red blood cell volume (a measure of anaemia), all of which are known to be affected by trypanosomiasis. The pedigree animals were genotyped for 170-200 microsatellite markers distributed throughout the genome to investigate associations between genetic markers and phenotypic traits. Analysis has revealed a number of postulated quantitative trait loci (QTL) conferring tolerance to trypanosomiasis. Three of the most promising QTL regions (detailed in **Section 3.1.1**) are analysed in this thesis.

CHAPTER 2

DETECTION OF SELECTION AT MICROSATELLITES

2.1 INTRODUCTION

2.1.1 Effects of selection on neutral markers

The detection of the effects of selection through examination of patterns of genetic diversity at linked, neutral markers is a research aim of growing importance. Adopting such an approach promises first to elucidate the relative contributions of natural selection and neutral drift to the evolutionary process. Second, it may help to infer map locations of genes with actions of medical or economic importance. Mapping such genes is usually achieved by analysis of coincidence of phenotype and marker polymorphism, either by pedigree-based linkage analysis or through detection of gametic disequilibrium at the population level. However, given the imprecision of mapping methods and the inherent difficulties in measurement of some phenotypes there is often a need for complementary methods for inferring gene location.

A number of theoretical predictions have been made concerning the population genetic behaviour of neutral loci in chromosome regions under selection. Here we test these predictions by studying a large number of mapped microsatellite loci in six diverse cattle populations of known phylogenetic relationship. The microsatellite loci studied divide into three classes; microsatellites known to be closely linked to genes, microsatellites in the MHC class II region on bovine chromosome 23, and finally microsatellites not known to be close to any genes.

In considering the effects of selection at a locus on linked markers, three main models of selection have been considered. Two models predicted to have similar effects on linked loci are positive selection for a favoured allele at a locus (hitch-hiking) (Smith and Haigh 1974), and selection against deleterious mutations at a locus (background selection)

(Charlesworth *et al.* 1993). In both models, selection reduces allelic diversity at neutral loci; through fixation of an allele at a linked locus associated with the favoured allele at the selected locus in the case of hitch-hiking and through elimination of alleles linked to eliminated deleterious alleles in the case of background selection (Charlesworth *et al.* 1993). Under the third model, overdominant (or 'balancing') selection, the effects on neutral linked loci are essentially the opposite. Selection for heterozygosity at a locus reduces the chance of allele loss through genetic drift, thus maintaining polymorphism and promoting retention of alleles at linked neutral loci (Gillespie 1997).

Retention of alleles is predicted to increase allelic coalescence time at both the selected and linked neutral loci (Hudson and Kaplan 1988; Kaplan *et al.* 1988). However, theory suggests that the effects may be short-range when compared with the effects of directional selection (Gillespie 1997; Hudson and Kaplan 1988). In addition to intra-population effects, selection should alter the inter-population evolutionary dynamics of linked neutral markers. Whereas hitchhiking at a locus may accelerate genetic differentiation between isolated populations, maintenance of polymorphism through overdominant selection should retard inter-population differentiation (Slatkin and Wiehe 1998).

2.1.2 The major histocompatibility complex and overdominant selection

The mammalian major histocompatibility complex (MHC) is a family of genes encoding cell-surface glycoprotein receptors involved in antigen-recognition and presentation to the T cells of the immune system (Klein 1986). The MHC genes divide into two classes (class I and class II) with different expression patterns and function, though both classes play an important role in defending against invading pathogens (Klein 1986).

Class I and II genes contain highly polymorphic regions encoding antigen-binding sites. In contrast with most genes, polymorphism at antigen-binding sites often predates speciation, with alleles persisting in separate lineages for many millions of years after divergence (Figueroa *et al.* 1994). Evidence suggests that this is due to balancing selection acting to maintain diversity (Hughes and Nei 1988; Hughes and Nei 1989). The nature of the balancing selection is not clear, but it is likely that heterozygotes recognise a greater range of pathogens than homozygotes, and so are more resistant to disease. Increased polymorphism at nearby neutral nucleotide positions has been observed (Hedrick *et al.* 1991), although it is unclear how far ranging this effect might be.

2.1.3 Bovine MHC (BOLA) organisation

The bovine MHC genes are located on chromosome 23, for which detailed genetic and physical maps are available (Band *et al.* 1998; Kappes *et al.* 1997). In contrast to humans and mice the bovine MHC genes are subdivided into two regions; the class IIa/class I region and the class IIb region (Andersson *et al.* 1988; van Eijk *et al.* 1995). The regions are separated by a large physical distance, probably due to chromosomal inversion (Band *et al.* 1998). The class IIa region contains the DQ and DR genes, which are very highly polymorphic in ruminants. The closely linked class I region also contains a number of highly polymorphic genes (Amills *et al.* 1998).

The class IIb region contains a number of novel class II genes exclusive to artiodactyla, and homologues of several human genes (Andersson *et al.* 1988). These, however, show limited polymorphism (Davies *et al.* 1994) and there is no evidence for their expression in lymphocytes (Stone and Muggli-Cockett 1993), suggesting that class IIa and class I genes are much more important than class IIb genes in antigen recognition. Accordingly, antigen-

driven balancing selection should be strongest in the MHC class IIa/class I region of the chromosome.

2.1.4 Models of evolution and application to microsatellite data

The infinite allele model (IAM) of mutation was proposed in 1964 to account for the levels of genetic variation (chiefly protein polymorphism at the time) observed in populations. The model assumes that the number of possible alleles at a locus is sufficiently large that mutation always gives rise to novel alleles (Kimura and Crow 1964). In the context of gene sequences, this is not unreasonable. For a sequence of n base pairs, there exist n^4 possible distinct alleles. For genes that are hundreds or thousands of nucleotides long, the number of hypothetically possible alleles is thus extremely large. In a finite population, the number of distinct alleles is necessarily limited. There is a balance between gain of alleles through mutation and loss through genetic drift (sampling of alleles from parental generation to form the next generation). Applying the assumptions of the IAM It is possible to predict the expected homozygosity (expected proportion of homozygotes under Hardy-Weinberg equilibrium) for a population when equilibrium between allele gain and loss is attained (Kimura and Crow 1964). Kimura and Crow (1964) showed that expected homozygosity is given by $F = \frac{1-2\mu}{4N_e\mu - 2\mu + 1} \approx \frac{1}{4N_e\mu + 1}$, where μ is mutation rate and N_e is effective population size. The shape of the allele distribution may also be predicted (Ewens 1972). Genetic distance measures have been developed for loci evolving in accordance with the IAM. These include Nei's standard genetic distance, D_S (Nei 1972), used in this study.

2.1.5 Microsatellites

Microsatellites are regions of DNA sequence consisting of arrays of tandem repeats of two to five nucleotides. They show much higher levels of polymorphism than other classes of

nucleotide sequence, with different alleles having different numbers of repeat units in the array. The high polymorphism is evidently due to a very high intrinsic mutation rate, estimated from studies of familial mutation events to be of the order of $1x10^{-3}$ per locus per gamete per generation (Weber and Wong 1993). The principal mode of evolution for microsatellites is believed to be slipped-strand mispairing during DNA replication and repair, leading to loss or gain of repeat units from the microsatellite array (Levinson and Gutman 1987). **Figure 2.1** illustrates how repeat units can be lost or gained when DNA strands fail to align perfectly during DNA replication as a consequence of the presence of a repetitive tract of sequence. The process by which slipped-strand mispairing causes loss or gain of repeats during DNA repair outside of DNA replication is very similar (Levinson and Gutman 1987).

The mode of microsatellite evolution illustrated in **Figure 2.1** clearly does not conform to a strict IAM model. Rather it is much closer to the stepwise mutation model (SMM) of Ohta and Kimura (1973). The SMM in its original form proposes that allelic states can be considered as integer values on a linear scale. Mutations, occurring with probability μ per generation, cause positive or negative unitary changes in allele state with equal likelihood (Ohta and Kimura 1973). Applied in the context of microsatellites, the model predicts that mutations cause gain or loss of single repeat units from the microsatellite array with equal probability. A consequence is that mutation frequently gives rise to alleles that are already represented in the population. In such a case, the novel mutant allele and its pre-existing counterpart are identical by state, but not by descent, a phenomenon termed 'homoplasy'.

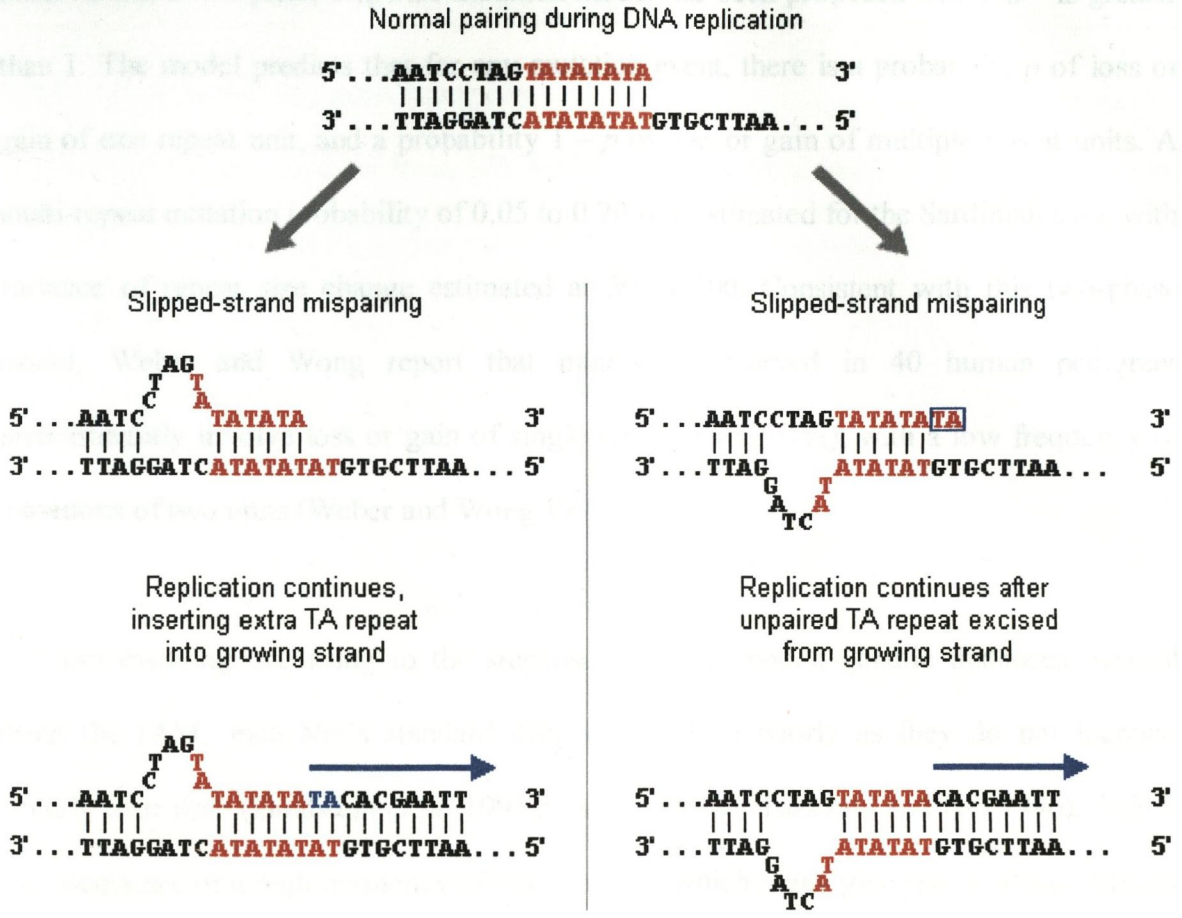

Figure 2.1: Slipped strand mispairing during DNA replication as a mechanism for microsatellite evolution

Blue arrows indicate starting point and direction of DNA synthesis. Vertical lines represent DNA base-pairing. A 2-base slippage occurs in a TA/AT repeat during replication of a DNA duplex, followed by strand elongation. Slippage (of the growing strand) in the 3'→5' direction (left side of figure) results in insertion of an extra TA repeat unit (coloured blue). Slippage in the other direction (right side of figure) results in loss of one TA repeat unit. This deletion results from excision of the unpaired repeat unit (bounded by blue box) at the end of the growing strand, presumably by the 3'→5' exonuclease activity of DNA polymerase. Modified from Levinson and Gutman (1987).

Under the generalised stepwise mutation model, mutations cause a change in allele repeat number of mean 0 and variance σ^2. For the strict stepwise mutation model, σ^2 is equal to 1, meaning that all mutations increase or decrease the microsatellite array of one repeat unit. A study of human genetic variation in Sardinia has found that the strict stepwise mutation model cannot account for the excessively high variance in microsatellite repeat number given the level of heterozygosity (Di Rienzo *et al.* 1994). To explain this

observation, a two-phase stepwise mutation model has been proposed where σ^2 is greater than 1. The model predicts that for any mutation event, there is a probability p of loss or gain of one repeat unit, and a probability $1 - p$ of loss or gain of multiple repeat units. A multi-repeat mutation probability of 0.05 to 0.20 was estimated for the Sardinian data, with variance of repeat size change estimated at 50 to 200. Consistent with this two-phase model, Weber and Wong report that mutations observed in 40 human pedigrees predominantly involve loss or gain of single repeat units (91%), with a low frequency of mutations of two units (Weber and Wong 1993).

For loci evolving according to the stepwise mutation model, genetic distances derived using the IAM, such Nei's standard distance, perform poorly as they do not increase linearly over time (Goldstein *et al.* 1995a; Slatkin 1995b; Takezaki and Nei 1996). This is a consequence of a high frequency of mutations, of which many give rise to alleles already represented in the population.

2.2 MATERIALS AND METHODS

2.2.1 Collection of cattle samples

DNA samples were obtained from 322 individual cattle from seven different populations from Africa, Europe and the Indian sub-continent (**Table 2.1** and **Sections 2.2.1a** to **2.2.1f**) Published data for three additional cattle populations was also used in the study (**Table 2.1**). These three populations are Hariana and Sahiwal (**Section 2.2.1b**) (MacHugh *et al.* 1997) and a second Holstein-Friesian population (**Section 2.2.1a**) (Schmid *et al.* 1999).

2.2.1a Charolais and Holstein-Friesian

For the Charolais and (Irish) Holstein-Friesian samples, blood or semen samples were obtained from artificial insemination (AI) stations in Ireland and Britain as described in MacHugh (1996). Supplementary semen samples were also received from the British AI service (now known as *Genus Ltd*). Pedigree records were consulted to ensure that there was the minimum degree of genealogical relatedness among the animals sampled. Published data for a second (Swiss) Holstein-Friesian population, described in (Schmid *et al.* 1999), was also used.

2.2.1b Indian zebu breeds

Hariana and Sahiwal samples were collected in northern India during April and May of 1990 at the National Institute for Animal Genetics (NIAG), Karnal in Haryana State, as described in MacHugh (1996). Sahiwal blood samples were taken from the large dairy herd kept at the adjoining National Dairy Research Institute (NDRI). Detailed pedigree records kept for these animals were consulted to avoid sampling related individuals. The Hariana animals were collected from a large university herd on the campus at Hissar Agricultural University. No information was available on their familial background.

Three additional purebred Indian Sahiwal bull samples were obtained from Dr. Pat Preston at the Centre for Tropical Veterinary Medicine, University of Edinburgh.

Ongole and samples were obtained from herds maintained at Visakhapatnam Government Dairy Farm, Andhra Pradesh. Nellore samples were obtained in Brazil. DNA was extracted and provided by C. Gaillard, Institut für Tierzucht, University of Berne, Switzerland.

2.2.1c Guinean N'Dama

Guinean N'Dama animals were sampled during the course of a field mission in February and March of 1992 as described in MacHugh (1996). The sampling location was approximately 350 km south of the Gambia within the tsetse zone. The 63 samples used were obtained from animals from 17 herds sampled in six locations including areas in and around the Fouta Djalon plateau, the region where the N'Dama breed was originally formed.

2.2.1d Djakoré

Djakoré were sampled during the course of a sampling trip in November 1996 undertaken in collaboration with the Institut Senegalais de Recherches Agricole (ISRA), Dakar. Samples were collected in two regions of Western Senegal; the Saloum Delta region north of the Gambia, and the Casamance region south of the Gambia. Nine herds were sampled in the Saloum delta region and ten in the Casamance. A total of 117 animals were sampled, including a large number of dam-offspring pairs. Only one member of any pair was used in this study to avoid typing related individuals.

2.2.1e *Togolese Somba*

Togolese Somba samples were obtained in collaboration with Dr Balabadi Dao. Samples were collected during October and November 1996. A sample of purebred animals was obtained from herds in four villages approximately 100km from the northern border with Burkina Faso. A sample of hybrid animals with a zebu genetic component was obtained from herds in five villages further north (from 10km to 60km from the Burkina border). DNA extractions were performed the following year by Dr Dan Bradley, using the protocol described in **Section 2.2.3**.

2.2.1f *Benin Somba*

The Benin Somba samples were obtained during the course of a sampling trip from October to December 1992 in collaboration with Prof. L.O. Ngere of the Dept of Animal Science, the University of Ibadan, Nigeria and with Dr. L. Gnaho, Benin.

2.2.2 Sampling strategy for African breeds

No pedigree information was available for any of the African breeds. Therefore to avoid sampling related individuals, samples were taken from as many locations as possible and if possible, no more than one tenth of a particular herd was sampled at any one location. In most instances, it was possible to ensure that the cattle sampled were unrelated by drawing upon the knowledge of the owners and herders of the animals. Blood samples were taken from many small village herds scattered around a particular region and transported in iceboxes back to the laboratory for DNA extraction.

2.2.3 DNA extraction from cattle samples

A salt extraction protocol was used for extraction of DNA from blood samples for the Togolese Somba samples and the Djakoré samples from Senegal. Extraction protocols used

for the European samples, Guinean N'Dama and Sahiwal and Hariana are described in MacHugh (1996).

2.2.3a Salt protocol for DNA extraction using fresh blood

1) Centrifuge 10 ml of whole fresh blood for 15 mins at 3,000 r.p.m.

2) Carefully remove the buffy coat (white cell) layer using a disposable Pasteur pipette and introduce into a clean 13 ml glass or plastic push-cap centrifuge tube.

3) With a 50 ml disposable plastic syringe, add approx. 5 ml of 50% (0.5x) Lysis Buffer [10mM $KHCO_3$, 0.155M NH_4Cl, 1mM EDTA (pH8.0)], taking care not to cross-contaminate the tubes with the syringe tip.

4) Invert the tube a few times and leave at room temperature for 15 mins.

5) Centrifuge for 15 mins at 3,000 r.p.m. Pour off the supernatant and gently rinse the cell pellet with distilled H2O, from a wash bottle (Take care not to cross-contaminate the tubes with the tip of the wash bottle.)

6) With a 50 ml disposable plastic syringe add 5 ml of Digestion Buffer [0.1M NaCl, 10mM Tris-HCl (pH8.0), 25mM EDTA (pH8.0), 20% SDS, 0.2mg/ml Proteinase K, stored at 4°C and warmed before use if necessary to redissolve an precipitate].

7) Break up the cell pellet with a large bore needle or pipette. Replace bung and incubate overnight at 50°C. (Place a weight on top if necessary to ensure that the bungs are not pushed out when the tubes are placed in the water bath.)

8) With a 50 ml disposable plastic syringe, add 5 ml of 6M NaCl. Replace the bung and shake vigorously for 15 seconds. Centrifuge for 15 mins at 3,000 r.p.m.

9) Using a disposable Pasteur pipette, remove the supernatant, avoiding the precipitated protein pellet and introduce into a new centrifuge tube.

10) Add 6 ml of room temperature absolute ethanol. Gently invert the tube until the DNA precipitates.

11) With a disposable plastic Pasteur pipette, fish out the DNA. Place into a clean, labelled 2 ml screw-top tube, with as little liquid carry-over as possible.

12) Add 2 ml absolute ethanol. Tightly fit the top and store at -20°C .

13) Associated salt can be removed from precipitated DNA by washing repeatedly with 70% ethanol.

2.2.3b Salt protocol for DNA extraction using frozen blood

1) Defrost blood.
2) Add an equal volume (5 ml) of 100% (1x) blood lysis buffer. Mix by inversion and leave for 15 minutes at room temperature.
3) Centrifuge at 3000 rpm for 15 minutes.
4) Pour off supernatant.
5) Add 5 ml of 50% (0.5x) blood lysis buffer, and proceed as for steps 4 to 12 for the fresh blood protocol.

2.2.4 Geographic origins and descriptions of breeds studied in Chapter 2

The breeds studied include two European *B. taurus* breeds (Charolais and Holstein-Friesian), two West African *B. taurus* breeds (N'Dama and Somba) and four Indian *B. indicus* breeds (Ongole, Nellore, Hariana and Sahiwal) (**Table 2.1**). For each of these, approximate breed origins have been identified. Of the four *B. indicus* breeds, the Sahiwal originated furthest north in the Punjab in Pakistan (**Figure 2.2**). Sahiwal are small to medium red animals selected for milk production in arid conditions (Felius 1995). The Hariana derive from the Haryana plains in Northern India (**Figure 2.2**). They are large animals selected principally for draught purposes, and thrive in semi-arid areas (Felius 1995). The Ongole breed originated further south on the western coast of India in Andhra Pradesh (**Figure 2.2**). The closely related Nellore breed was created from cattle transported from this region to Brazil starting in 1895 (Felius 1995). Ongole cattle are large animals bred as dual-purpose dairy and draught animals. The derived Nellore cattle are heavier and have been selected for beef production (Felius 1995).

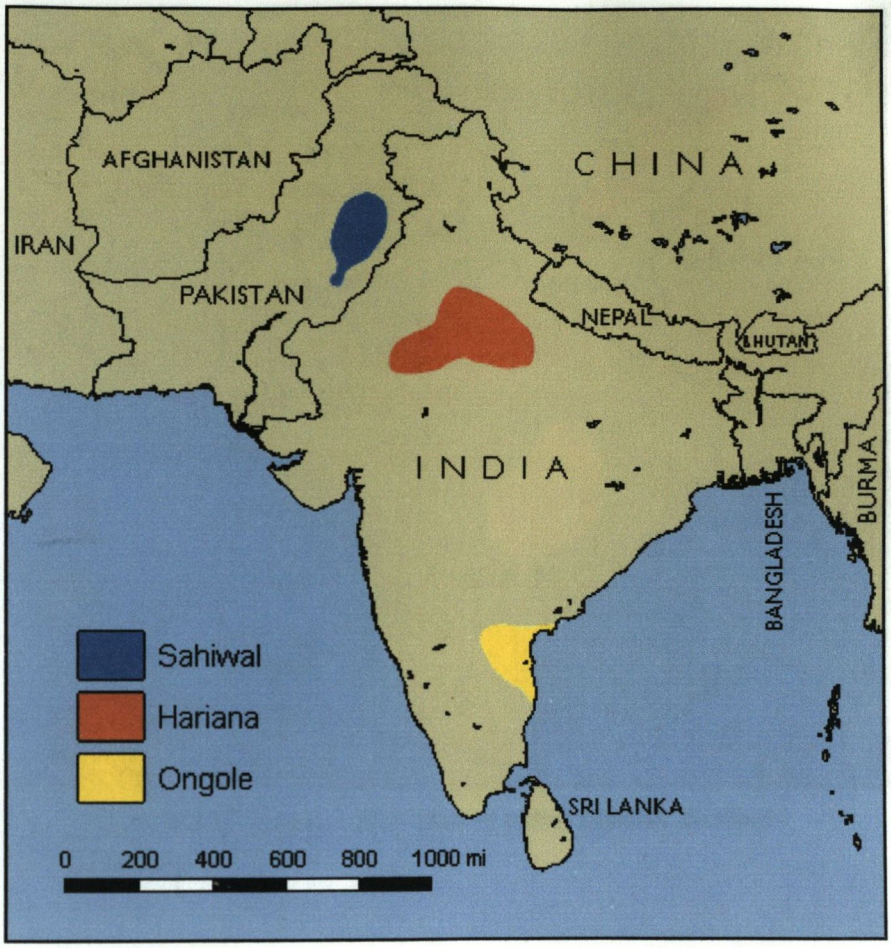

Figure 2.2: Sites of origin of *Bos indicus* (zebu) cattle breeds studied
Adapted from Felius (1995)

Holstein-Friesian cattle originated in the Netherlands (**Figure 2.3**), although cattle introduced from Denmark during the 18th Century have also contributed to the breed. They are large black and white animals that have been selected to give very high milk yields. This has led to the export of Holstein-Friesian cattle to many countries outside the Netherlands (Felius 1995). The Charolais breed originated in north-central France (**Figure 2.3**). Charolais cattle are large creamy-white animals selected primarily for beef (Felius 1995).

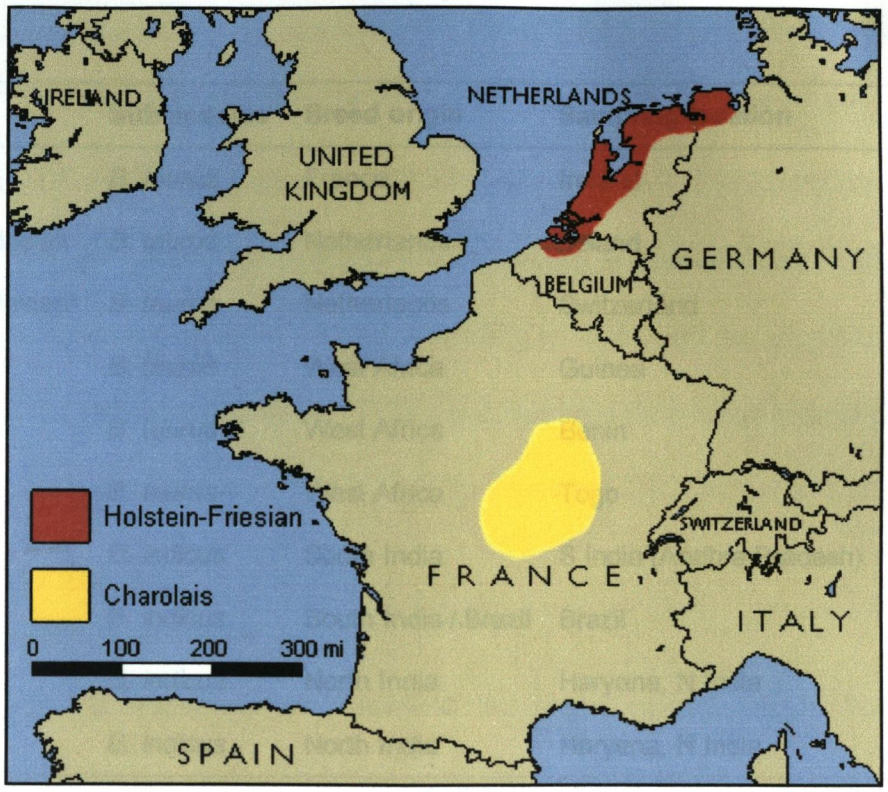

Figure 2.3: Sites of origin of European *Bos taurus* cattle breeds studied
Adapted from Felius (1995)

N'Dama cattle originated in West Africa in and around Guinea, the source of the samples used in this study (**Figure 1.2**). N'Dama cattle are longhorn animals that are small in stature (approximately two thirds the height of most European *B. taurus* cattle) and variable in colour. They are multi-purpose animals used for milk, draught and beef (Felius 1995). A defining characteristic of N'Dama is innate tolerance to tsetse-borne trypanosomiasis (Trail *et al.* 1979). Animals can remain productive in areas of trypanosomiasis challenge where other breeds would succumb to disease. Somba cattle also derive from West Africa, coming from a region centring on Benin to the east of the source of N'Dama cattle (**Figure 1.2**). Unlike N'Dama, Somba are shorthorn animals. In common with N'Dama, though, Somba are small animals (~90-110cm at withers) (Trail *et al.* 1979) showing tolerance to trypanosomiasis, although to a slightly lesser degree than N'Dama (Murray *et al.* 1982). Two separate populations of Somba were sampled; the first in Benin and the second in neighbouring Togo (**Figure 1.2**).

Breed	Sub-species	Breed origin	Sampling location	Sample size
Charolais	*B. taurus*	France	Ireland	34
Holstein-Friesian	*B. taurus*	Netherlands	Ireland	40
Holstein-Friesian*	*B. taurus*	Netherlands	Switzerland	50
N'Dama	*B. taurus*	West Africa	Guinea	63
Somba	*B. taurus*	West Africa	Benin	60
Somba	*B. taurus*	West Africa	Togo	60
Ongole	*B. indicus*	South India	S India (Andhra Pradesh)	30
Nellore	*B. indicus*	South India / Brazil	Brazil	35
Hariana†	*B. indicus*	North India	Haryana, N India	10
Sahiwal†	*B. indicus*	North India	Haryana, N India	13

Table 2.1: Details of cattle breeds studied

* Data taken from Schmid *et al.* (1999)

† Data taken from MacHugh *et al.* (1997)

2.2.5 Microsatellite loci studied

Sixty-seven different polymorphic microsatellite loci distributed throughout the bovine genome were selected for study. For 20 loci, genotypic data was already available for a number of cattle populations (Loftus *et al.* 1999, Schmid, 1999 #11; MacHugh *et al.* 1997). For each of the remaining 48 loci, six cattle populations were genotyped: the two zebu populations (Ongole and Nellore), the two European populations (Charolais and Irish Holstein-Friesian), the Guinean N'Dama population and either the Togolese or the Benin Somba population.

2.2.6 PCR amplification of microsatellite loci

2.2.6a *Non-radio-labelled PCR*

Optimisation PCR reactions were carried out in 125µl polypropylene tubes or in 96-well 125µl polycarbonate microtitre plates, depending on the number of reactions to be

performed. Reactions included: reaction buffer (50mM KCl, 10mM Tris-HCl pH 9.0, 0.1% Triton X-100); 200µM dATP, 200µM dGTP, 200µM dTTP, 200µM dCTP; 0.2µM each PCR primer; 0.5 units Taq polymerase; 20-30ng sample DNA. Total reaction volume was 11.0µl, with a 10µl mineral oil overlay. Thermocycling was conducted: 4 min at 95°, 35 cycles of 30 sec denaturing at 94°, 40 sec annealing at 52-60°, 40 sec extension at 72°, and a final extension step of 5 min at 72°. PCR yield and specificity were assessed by running products out on 1-2% agarose gels with a 100bp increment DNA ladder as a size standard.

2.2.6b Radio-labelled PCR

PCR reactions for typing of microsatellite loci were carried out in 96-well 125µl polycarbonate microtitre plates. Reaction components were as for optimisation reactions, except the dCTP concentration was reduced to 10µM and 0.14µCi 32-P dCTP was included for radiolabelling of PCR product. Total reaction volume was 5.5µl, with a 10µl mineral oil overlay.

2.2.7 Visualisation of PCR products on polyacrylamide gels

Samples were mixed with 5.0µl formamide loading buffer (98% deionised formamide, 10mM EDTA (pH 8.0), 0.025% Bromophenol blue 0.025%Xylene cyanol FF), heat-denatured by holding at 94°C for 3 minutes, and snap-cooled by placing on ice to prevent renaturing. 3µl of each sample was loaded into a 96-lane 6% polyacrylamide sequencing gel. Gels were prepared using the Sequagel^{TM} system (National Diagnostics). Following the manufacturer's instructions, sequencing gels were prepared containing 6% Acrylamide:Bis Acrylamide (19:1), 50% urea, 100mM Tris-borate (pH 8.3) and 2mM EDTA (pH 8.0). Gels were polymerised by addition of 400 µl of 10% APS and 40 µl of TEMED solutions prior to pouring. Gels were run using 1x TBE electrophoresis buffer (0.1M Tris-borate (pH 8.3), 0.002M EDTA (pH 8.0)). Standardisation across gels was

performed by the inclusion of PCR products from the same animals on each gel. Gels were electrophoresed on a vertical sequencing gel electrophoresis apparatus at approximately 70W for 2-4 hours depending on the size of the PCR product. After running, gels were applied to a sheet of Whatman filter paper, covered with clingfilm and dried at 80 °C for 1 hour on a Savant SGD 4050 vacuum slab gel drier. The dried gels were exposed to Agfa autoradiographic X-ray film in a film cassette for 6 to 48 hours, depending on the intensity of the signal. The exposed autoradiograms were developed using an Agfa Curix film developer.

2.2.8 Analysis

A number of different population-genetic analyses were performed using a variety of different software. Formatting of data for the various software listed was performed using 'the Excel Microsatellite Toolkit' described in **Section 6.1**.

2.2.9 Hardy-Weinberg equilibrium

All population-locus combinations were tested for significant deviation from Hardy-Weinberg equilibrium (HWE), or random association of gametes. The test performed is based on the use of Fisher's exact test for Hardy Weinberg proportions (Louis and Dempster 1987). The exact test is performed by generating all possible genotypic $k * k$ contingency tables (where k is the number of haplotypes) with the same marginal counts as the observed table, and calculating their probabilities. Tables with more extreme genotypic distributions (i.e. more extreme departures from HWE proportions) have correspondingly lower probabilities. The exact test probability is obtained by summing the probabilities of tables whose probabilities are equal to or less than that of the observed distribution. The exact test can be fully enumerated if the number of haplotypes is small. However, construction of all possible contingency tables rapidly becomes very computationally

intensive as the number of distinct haplotypes increases. In this event, a Markov chain can be used to explore the space of contingency tables with the same marginal counts as the observed table by randomly generating large numbers of genotype tables and calculating the proportion whose probability is less than or equal to that of the observed genotypic distribution (Guo and Thompson 1992).

Deviation from HWE was tested using GenePop (Raymond and Rousset 1995). For loci with fewer than five haplotypes, all contingency table probabilities were calculated. For loci with five haplotypes or more, the Markov chain method was used with 10,000 dememorisation steps to move from the initial, observed genotype distribution to a random starting point. 200 batches of 1000 iterations were then performed to estimate the probability of obtaining the observed genotypic table, as described, with probabilities averaged across batches (Guo and Thompson 1992).

2.2.10 Number of distinct alleles per locus

To obtain an indication of the relative genetic diversity at different loci, the number of distinct alleles at each locus was calculated. This quantity varies considerably with sample size (Nei 1987). There was considerable variation in the sample size for some of the populations used in this study. To correct for this, a resampling approach was used to standardise sample sizes for all population-locus combinations. After excluding the zebu breeds, due to very small sample sizes for two of the breeds, the smallest number of chromosomes (n_{min}) typed at any locus was determined. For all population-locus combinations, n_{min} alleles were then randomly sampled, with replacement, and the number of distinct alleles drawn was obtained. Results of 100 resampling replicates were averaged to give a final estimate.

2.2.11 Gene diversity

Nei's unbiased gene diversity is better indicator of genetic diversity as it is less sensitive to sample size than the number of distinct alleles at a locus. For diploid organisms this is equivalent to expected heterozygosity, which is the proportion of heterozygotes expected if the population is in Hardy-Weinberg equilibrium (Nei 1987). For a sample of n individuals taken from a population in which the frequency at a locus of genotype A_iA_j is denoted $\hat{X}_{ij}$, Nei's unbiased gene diversity is given by the formula:

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

Equation 2.1

where

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

Equation 2.2

Average gene diversity (across r loci) is given by:

$$\hat{H} = \sum_{j=1}^{r} \hat{h}_j / r$$

Equation 2.3

Variance of gene diversity, which takes into account the effects of sampling genes from a population and loci from the genome, is given by the formula:

$$V(\hat{H}) = V(\hat{h})/r$$

Equation 2.4

where

$$V(\hat{h}) = \sum_{j=1}^{r} \left(\hat{h}_j - \hat{H}\right)^2 /(r-1)$$

Equation 2.5

Gene diversity estimates were obtained using 'the Excel Microsatellite Toolkit' described in **Section 6.1**.

2.2.12 Comparison of observed and neutral expectation allele distributions

Ewens has shown that the null allele frequency distribution under mutation-drift equilibrium (MDE) can be estimated if the sample size and number of distinct alleles is known (Ewens 1972). Comparing the observed allele distribution with the corresponding estimate can reveal departure from MDE, possibly due to selection.

For each polymorphic population-locus combination, Watterson's statistic F, equal to the expected homozygosity, was calculated according to the formula:

$$F = \sum_{1}^{k} x_i^2$$

Equation 2.6

where k is the number of distinct alleles in the population-sample and x_i is the frequency of the i'th allele in the sample (Watterson 1978). The observed F value (F_{obs}) was compared with the corresponding estimated null F values (F_{exp}) obtained under the assumption of either the infinite allele mutation model (IAM) or the strict stepwise-mutation model (SMM). Null F_{exp} values were estimated using the program Bottleneck (Cornuet and Luikart 1996) which uses the coalescent process (Hudson and Kaplan 1988) to simulate MDE genealogies with the same number of alleles and distinct haplotypes as in the observed sample. F_{exp} is calculated and averaged across genealogies, and can be compared with F_{obs}.

2.2.13 Allele frequency distributions for simulated hybrid populations

To test the effects of population-hybridisation on the relationship between observed and predicted (null) F values, simulated hybrid populations were constructed using data described in this study for 49 anonymous microsatellite loci typed in Charolais, Holstein-Friesian, N'Dama, Somba, Nellore and Ongole populations. The two European *B. taurus* breeds were pooled, as were the two African *B. taurus* breeds and the two Indian *B. indicus*

breeds to form three parental populations. Hybrid populations were constructed using pairwise combinations of these three parental populations. For each pairwise combination, the relative genetic contribution of the two parental populations to the simulated hybrid population was varied in steps between 0% and 100%. At each step, 50 replicate populations were constructed by randomly sampling (with replacement) 80 alleles from the parental populations, in accordance with the relative genetic contributions from each. This sampling was performed for each of the 49 loci used. For each population-locus combination in each replicate population, F_{obs} was calculated, and compared with the corresponding null value F_{exp} estimated under Ewen's sampling theory (Ewens 1972). Null F values were calculated using the program Arlequin (Schneider *et al.* 1997) by simulating, under Ewens' sampling theory, 1000 random samples with the same number of individuals and the same number of distinct alleles as the observed sample, using the algorithm of Stewart (Fuerst *et al.* 1977).

2.2.14 Nei's standard genetic distance

Nei's standard genetic distance (D_S), proportional to the mean number of codon substitutions per locus (Nei 1972), was calculated for all population-locus combinations. For a pair of populations X and Y, the frequency of allele A_i in population X is denoted by x_i and in population Y by y_i. If one allele is chosen randomly from each population, the probability that the alleles will be identical is given by $j_{XY} = \sum x_i y_i$. For alleles drawn at random from the same population, the probability of identity is $j_X = \sum x_i^2$ for population X, and $j_Y = \sum y_i^2$ for population Y. The standard genetic distance between the two populations for a given locus can then be expressed as:

$$d_S = -\ln\left(j_{XY} / (j_X j_Y)^{1/2}\right)$$

Equation 2.7

Averaging over all loci, Nei's standard distance is given by:

$$D_S = -\ln\left(J_{XY}/(J_X J_Y)^{1/2}\right)$$

Equation 2.8

where J_{XY}, J_X and J_Y are the averages across all loci of j_{XY}, j_X and j_Y. Under the infinite allele model of mutation, where mutation always leads to a new allele, and with the additional condition that loci are in MDE, D_S is expected to increase linearly with evolutionary time (Nei 1972; Takezaki and Nei 1996). D_S was calculated using the program Microsat v1.5 (Minch 1995).

2.2.15 Delta mu squared distance

Delta mu squared is a genetic distance derived for loci evolving under the stepwise mutation model in populations at MDE. Under these conditions, the variance in microsatellite repeat number within each population is expected to be constant over time, whereas the difference in the mean number of repeats between populations increases linearly. The genetic distance delta mu squared exploits this:

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

Equation 2.9

where μ_A and μ_B are the mean numbers of repeats for alleles in populations A and B respectively (Goldstein *et al.* 1995b). The distance has been shown to increase linearly with time, and to have a lower variance than other distances derived under the stepwise mutation model (Goldstein *et al.* 1995b). $(\delta\mu)^2$ was calculated using the program Microsat v1.5 (Minch 1995).

2.3 RESULTS

67 microsatellite loci were studied in a range of cattle populations to determine whether genomic environment influenced the population genetics of the loci.

2.3.1 Microsatellites studied

The microsatellites studied are all predominantly dinucleotide repeats. Published sequences, available for 58 of the 67 loci studied (**Table 2.2**), show that one locus (ETH10) contains a significant trinucleotide repeat, and another (DRB3) contains a hexanucleotide repeat, but that both also contain stretches of dinucleotide repeat. Repeat sequences for the remaining 56 microsatellites sequenced all consist of perfect or imperfect dinucleotide repeats. 54 of the 58 loci sequenced contain tracts of CA/GT repeats. About half of the microsatellites sequenced are entirely or almost entirely pure repeats (unbroken tandem repeats of the same dinucleotide unit – e.g. BMS468, TAMLS113.3). The rest are interrupted repeats (e.g. PRL) and compounds of different arrays of different repeat units (e.g. BoLA DRB1).

For 59 of the 67 loci, the alleles observed in the cattle populations typed in this study are consistent with a mode of mutation involving loss or gain of integer numbers of dinucleotide repeat units. The other eight loci have alleles that differ from one another by non-integer numbers of dinucleotide repeats.

2.3.2 Microsatellite classification

The microsatellite loci surveyed were each assigned to one of three classes: MHC-linked, gene-linked or anonymous (**Table 2.2**). Loci classed as MHC-linked are those five closest to the MHC class IIa/class I region in the central part of chromosome 23. Other loci on

chromosome 23, including those near to the class IIb region, are not classed as MHC-linked given there is no *a priori* expectation of strong balancing selection associated with antigen recognition in the vicinity of these loci. Eight microsatellites in close proximity to known bovine genes are classed as gene-linked. Of these, HBB, PRL and RASA each contain a tandem repeat in the 5' untranslated region. OCAM, RBP3 and HMH1R each contain a microsatellite repeat in the 3' untranslated region. Locus TGLA116 was found to be linked to the progressive degenerative myeloencephalopathy gene (PDME, or 'weaver') at a recombination frequency of 3% in a pedigree of Brown Swiss cattle (Georges *et al.* 1993). Finally, the IL4 gene has a microsatellite repeat approximately 35 kilobases upstream (Buitkamp *et al.* 1995). The remaining 54 loci are not known to be closely linked to genes, and are classed as anonymous.

Of the ten cattle population samples used, two (Charolais and N'Dama) were typed at all 67 microsatellite loci studied. The remaining eight populations were typed at a subset of the 67 loci (**Table 2.3**).

Details of 67 microsatellite loci surveyed

Locus	Classification Description	Chr	Position (cM) *	GenBank No	Forward primer Reverse primer	Repeat motif †	References
BMS468	MHC	23	36	G18838	GTT AAG CAG AGG GTT TCC CC TAT TCC CAG GTG CTC TGA GG	$(AC)_{11}$	16, 31
BoLA DRB1	MHC MHC class II DRß pseudogene	23	35.4	M30010	ATG GTG CAG CAG CAA GGT GAG CA GGG ACT CAG TCT CTC TAT CTC TTT G	$(GT)_{18}(GA)_8$	16
CYP21	MHC Cytochrome P450 subfamily XXI	23	36	M11267	GGA GGG TTA CAG TCC ATG AGT TTG TCG CGA TCC AAC TCC TCC TGA AG	$(CA)_{20}(TA)_2CATA(GA)_2(CA)_3$	9, 16
DRB3	MHC	23	35.7	S66481	GAG AGT TTC ACT GTG CAG CGC GAA TTC CCA GAG TGA GTG AAG TAT CT	$(TG)_{20}(AG)_8(ACAGAG)_4AT-(AGAC)_2AGAACA(GA)_2$	11
TAMLS113.3	MHC	23	38	L29386	TTA CTG CTG AGC CAC CGG GAT GGG GGT CAC AAA CTG AC	$(TG)_{18}(TA)_2$	16, 28
HBB	Gene ß-globin locus	15	39.6	M63453	GGG ACT CAT AGA CCA TTC ATA GC CAA CTG GCA AAA TTT CAT TCT T	$(TA)_2(CA)_{17}$	14, 16
HMH1R	Gene Histamine H1 receptor	22	-	E04992	GGC TTC AAC TCA CTG TAA CAC ATT TTC TTC AAG TAT CAC CTC TGT GGC C	$(TA)_2(CA)_{15}$	25
IL4	Gene Interleukin 4	7	30.5	-	GTG CTG GAC ATC TGC AAG TG ACA TTC AGG TCT GTG ATC CAT G	$[(T/G)A]_m(CA)_n$	7, 16
OCAM	Gene Opioid binding molecule	29	39.9	X12672	CCT GAC TAT AAT GTA CAG ATC CCT C GCA GAA TGA CTA GGA AGG ATG GCA	$(CT)_5(N)_8(CT)_4(GA)(CT)_3(N)_7(CT)_8$	16, 24
PRL	Gene Prolactin gene	23	43.2	X16641	GGA AAG TGA ACA TGA CTG TCT AG GCC CTC TCT TCT ACA ATG AAC AC	$(TG)_6(TA)(TG)_6$	9, 16
RASA	Gene RAS p21 protein activator	7	103.1	X12602	CCC TTC CGC TTT AGT GCA GCC AG GGG CCA CAG CCC AGG ATC GGG AGC	$(TG)_{12}$	10, 16
RBP3	Gene Retinol-binding protein 3	28	52.4	M19686	GAC CTT CTA TGC TTC CAC TCT AG GCT TTA GGT AAT CAT CAG ATA GC	$(CT)_2(TC)_6(ACAT)_2A(CA)_{15}(TA)_2-(N)_6(AT)_4(TA)_5$	13, 16
TGLA116	Gene Weaver gene (PDME)	4	48.9	-	GCA CAG TAA TAA GAG TGA TGG CAG A TGG AGA AGA TTT GGC TGT GTA CCC A	$(TG)_n$	8, 15, 16, 37
BL1001	Anonymous	2	43.7	-	AGA CGA GGC AAC TTG GAA TCT CGT GTC AGA AAA CAT AAC TGC C	-	16, 29
BL1067	Anonymous	7	14.2	-	AGC CAG TTT CTT CAA ATC AAC C ATG GTT CCG CAG AGA AAC AG	-	16, 29

Locus	Classification Description	Chr	Position (cM) *	GenBank No	Forward primer Reverse primer	Repeat motif †	References
BM1258	Anonymous	23	23.9	G18385	GTA TGT ATT TTT CCC ACC CTG C GAG TCA GAC ATG ACT GAG CCT G	$(GT)_{16}$	2, 16
BM1815	Anonymous	23	19.8	G18389	AGA GGA TGA TGG CCT CCT G CAA GGA GAC AAG TCA AGT TCC C	$(GA)_{17}(TG)_8$	16, 27
BM1818	Anonymous	23	50.9	G18391	AGC TGG GAA TAT AAC CAA AGG AGT GCT TTC AAG GTC CAT GC	$(TG)_{13}$	2, 16
BM1824	Anonymous	1	108.6	G18394	GAG CAA GGT GTT TTT CCA ATC CAT TCT CCA ACT GCT TCC TTG	$(GT)_{15}$	2, 16
BM2113	Anonymous	2	106.2	M97162	GCT GCC TTC TAC CAA ATA CCC CTT CCT GAG AGA AGC AAC ACC	$(CA)_{20}$	2, 16
BM3010	Anonymous	2	35.3	G18785	TCA TCT TTG TCA AGA CCT GGC AGT GGG AGA GGG CTT TGG	$(CA)_4TA(CA)_{19}$	16, 31
BM6105	Anonymous	7	35.7	G18589	ACT AAT AAG AAA TTC TGC ATG TGT G CCA CCA TGA CTC AGA AGT AGT TC	$(TG)_7TA(TG)_{14}$	16, 31
BM7233	Anonymous	23	49.1	G18795	GGA ATG AAA GAG CCT AGC AGC AGG ACT ACT GTA TGG ATG TGC G	$(TG)_{18}$	16, 31
BM733	Anonymous	5	113.5	G18449	ATG CTC CTT GTG CTC TCA CC GCC ATA GGA GAA AAA CTT GGG	$(GT)_{14}$	2
BMS1116	Anonymous	7	30.5	G18618	GAG CTT CGA GAA GGT TGG TG TCT GTG TGC ATG TCT GCG T	$(CA)_{14}CG(A)_3(CA)_4$	16, 31
BMS1300	Anonymous	2	46.6	G18651	TCC TGG GCC TGG TAA AAT AT CTT TTG GTT TGG AAA GAA GTT G	$(TA)_2(CA)_4GA(CA)_3TA(CA)_{13}$	16, 31
BMS2024	Anonymous	2	55	G18715	TAT AGC CTT GCT GTA AGA ATT GTG TGC TCT AAA GGC CAG TGT AAT AA	$(TG)_{11}AG(TG)_{10}$	16, 31
BMS2269	Anonymous	23	60.8	G18936	AAT TCA CCC ACT GCC AGC TGA AAT ATG CTT ACC TCC CAC C	$(GT)_{14}TT(GT)_5$	16, 31
BMS597	Anonymous	5	120	G18887	AGA AGA AGA GCA AGA GCA AAG G AAG AAG ACC TCT CTG ACC CTC C	$(TG)_{15}$	31
BMS713	Anonymous	7	15.2	G18886	CCA AGG GAG GAA AAA TAA GTT AA ACC AGC AGT AGG TTG AGG TTA A	$(CA)_{19}$	16, 31

Locus	Classification Description	Chr	Position (cM) *	GenBank No	Forward primer Reverse primer	Repeat motif †	References
BMS772	Anonymous	5	106.2	G18877	TTG TGC AAT CAA GTG GTA ACT G CTC ACT AAG ATG CCT GGT GAT C	$(TG)_{14}$	31
BMS803	Anonymous	2	41	G18580	GAG GTA GGG AAT CAG GTA AGG C AGC TGC ATG GCT GAA CAA G	$(CA)_{19}$	16, 31
BMS8126	Anonymous	5	122.1	G18854	TCT TCC TCC GCA GAC TGG GAA CCT GTG GAT GAG CGG	$(TG)_2AG(TG)_3AG(TG)_{12}$	31
BP34	Anonymous	23	57.5	-	ATT TGA AAA GGC CTG TGA GG TAG CTT GAC TCC AAC CTC TTC C	-	2, 16
CSRM60	Anonymous	10	-	-	AAG ATG TGA TCC AAG AGA GAG GCA AGG ACC AGA TCG TGA AAG GCA TAG	-	1
CSSM005	Anonymous	23	7.2	U03785	TGT ACT ATA TAA GCA CCA GAG AGT GTC CTT ACT ATC TTT AAG TGA CTG	$(TG)_{18}$	16, 26
CSSM024	Anonymous	23	58.4	U03812	GAG GAT GCA GAA CCA GGC GTT AGA GGT CGA AGA GAG TCA GAC ATG ACT	$(GT)_2(GA)_2(GT)_4(GA)_2(GT)_4(GA)_2(GT)_9$- $(GA)_2(GT)_3(GA)_2(GT)_3(GA)_2(GT)_{15}$	16, 26
CSSM042	Anonymous	2	34.4	U03818	GGG AAG GTC CTA ACT ATG GTT GAG ACC CTC ACT TCT AAC TGC ATT GGA	$(GT)_{17}(CT)_2(GT)_7CT(GT)_5CT(GT)_4$	16, 26
CSSM066	Anonymous	14	-	-	ACA CAA ATC CTT TCT GCC AGC TGA AAT TTA ATG CAC TGA GGA GCT TGG	-	1
ETH10	Anonymous	5	70	Z22739	GTT CAG GAC TGG CCC TGC TAA CA CCT CCA GCC CAC TTT CTC TTC TC	$(TC)_2CC(TC)_3(N)_{70}$- $(CCT)_2(CCA)_5(A)_3(AC)$	34
ETH131	Anonymous	21	32.3	Z14039	GTG GAC TAT AGA CCA TAA GGT C GCT GTG ATG GTC TAC GAA TGA	$(CG)_5(C)(CA)_{19}$	16, 30
ETH152	Anonymous	5	118.3	Z14040	TAC TCG TAG CGC AGG CTG CCT G GAG ACC TCA GGG TTG GTG ATC AG	$(CA)_{17}$	30
ETH2	Anonymous	5	108.5	Z22743	CCC ACA GGT GCT GGC ATG GCC CCA TGG GAT TTG CCC TGC TAG CT	$(CA)_{16}$	34
ETH225	Anonymous	9	8.1	Z14043	GAT CAC CTT GCC ACT ATT TCC T ACA TGA CAG CCA GCT GCT ACT	$(TG)_4(N)_4(CA)_{18}$	16, 30
HEL1	Anonymous	15	27.7	X65202	CAA CAG CTA TTT AAC AAG GA AGG CTA CAG TCC ATG GGA TT	$(CA)_{16}$	16, 17

Locus	Classification Description	Chr	Position (cM) *	GenBank No	Forward primer Reverse primer	Repeat motif †	References
Hel13	Anonymous	11	114.5	X65207	TAA GGA CTT GAG ATA AGG AG CCA TCT ACC TCC ATC TTA AC	$(AG)_4(AC)_5AT(AC)_{18}$	16, 17
HEL5	Anonymous	21	13	X65204	GCA GGA TCA CTT GTT AGG GA AGA CGT TAG TGT ACA TTA AC	$(CA)_{22}$	16, 17
HEL9	Anonymous	8	76.7	X65214	CCC ATT CAG TCT TCA GAG GT CAC ATC CAT GTT CTC ACC AC	$(TG)_3CAGA(TG)_{24}$	16, 17
ILSTS001	Anonymous	7	16.6	L14438	GGT GCT GTT ATC TAG AAT TTG G GGA GTC ATA CAC AAC TGA GC	$(TG)_{15}$	3, 16
ILSTS005	Anonymous	10	97.4	L23481	GGA AGC AAT GAA ATC TAT AGC C TGT TCT GTG AGT TTG TAA GC	$(TG)_4(TACA)(TA)_5(TG)_9(TA)(TG)_2(TA)_7$	5, 16
ILSTS006	Anonymous	7	116	L23482	TGT CTG TAT TTC TGC TGT GG ACA CGG AAG CGA TCT AAA CG	$(GT)_{23}$	6, 16
ILSTS014	Anonymous	19	49.1	L23488	CTG ACT ATG GTG ATA ATC CC TCT TTT CCC TTT CCT TCC CC	$(TG)_{10}$	4, 16
ILSTS030	Anonymous	2	35.3	L37212	CTG CAG TTC TGC ATA TGT GG CTT AGA CAA CAG GGG TTT GG	$(GT)_{10}AT(GT)_2$	16, 18
INRA005	Anonymous	12	83.1	X63793	CCT TTC AAA AAC ACG GAA ATT CGG GGG CTT CAG GCA TAC CCT ACA CCA CAT G	$(GT)_{13}$	16, 36
INRA023	Anonymous	3	-	X67830	GAG TAG AGC TAC AAG ATA AAC TTC TAA CTA CAG GGT GTT AGA TGA ACT C	$(CA)_{21}$	35
INRA032	Anonymous	11	-	X67823	AAA CTG TAT TCT CTA ATA GCT AC GCA AGA CAT ATC TCC ATT CCT TT	$(CA)_3(TA)(CA)_4(CG)(CA)_2(N)_{12}$- $(CA)_8(TA)(CA)_{13}$	35
INRA063	Anonymous	18	48.7	X71507	ATT TGC ACA AGC TAA ATC TAA CC AAA CCA CAG AAA TGC TTG GAA G	$(AC)_{13}(AT)_3$	16, 35
MGTG7	Anonymous	23	46.5	-	TTC ATT GCA GCA CTA TTT ACA ATA G TAA GTT CCC TGT ATC ATT TTT TGA	-	15
RM006	Anonymous	7	22.1	U32911	TAC AAT ATC TGG TCA CTG GA GAT CAC CAT ATT TAT GAG ATG	$(CA)_{13}$	16, 20
RM012	Anonymous	7	7.9	U03047	CTG AGC TCA GGG GTT TTT GCT ACT GGG AAC CAA GGA CTG TCA	$(CA)_3TGTA(CA)_{10}$	16, 21

Locus	Classification Description	Chr	Position (cM) *	GenBank No	Forward primer Reverse primer	Repeat motif †	References
RM033	Anonymous	23	17.3	U03054	GCT CAT TCT CCT GGG ATG CAG A GCT CCT TTA GTT TTC TTG TGG GAG	$(CA)_{14}$	16, 19
RM185	Anonymous	23	45.1	-	TGG CCT GCT TAT GCT TGC ATC GAG TTT CCT TTG CAT GCC AGT C	-	1, 16
RM356	Anonymous	2	51.9	G29110	GCA TCA CTA ACA TCC ACT GAG G CCA CTA GGA GAG GTC ATT CCC	$(GA)_2TA(CA)_{14}$-94bp-$(CA)_5$	16, 22
TGLA303	Anonymous	7	38.5	-	CTT GTG TGC CAG ACC CAG GAA TCC CAT AAG TCA AAG TAA CAG TTT AGA TGT CC	-	8, 15, 16
TGLA377	Anonymous	2	27.2	-	GAC TGT CAT TAT CTT CCA GCG GAG AGA CTT TGG ATC TCT GGT TGA AAT G	-	8, 15, 16
TGLA48	Anonymous	7	22.1	-	AAA TGT TTT ATC TTG ACT ACT AAG C ACA TGA CTC TGC CAT AGA GCA T	$(TG)_n$	13
UWCA1	Anonymous	23	22.1	L06454	AGA GTG TCT TAT AAT TAG CCA GGA AG AAC TCT TTC AGT TGG TTC CTG T	$(CA)_8C(CA)_{17}$	16, 32, 33

Table 2.2: Details of 67 microsatellite loci surveyed

'Chr' is the chromosome on which the locus is located. *Position within the chromosome is taken from the on-line MARC bovine genome map http://sol.marc.usda.gov/genome/cattle/cattle.html, originally published as Kappes *et al.* (1997)

† Repeat motif is that of the published (GenBank) allele. A dash ('-') indicates that repeat motif for a locus is unknown.

References: 1 - Barendse, Armitage *et al.* 1994; 2 - Bishop, Kappes *et al.* 1994; 3 - Brezinsky, Kemp *et al.* 1992; 4 - Brezinsky, Kemp *et al.* 1993; 5 - Brezinsky, Kemp *et al.* 1993; 6 - Brezinsky, Kemp *et al.* 1993; 7 - Buitkamp, Schwaiger *et al.* 1995; 8 - Crawford, Dodds *et al.* 1995; 9 - Creighton, Eggen *et al.* 1992; 10 - Eggen, Solinas-Toldo *et al.* 1992; 11 - Ellegren, Davies *et al.* 1993; 12 - Ferretti, Leone *et al.* 1994; 13 - Fries, Hediger *et al.* 1988; 14 - Fries, Eggen *et al.* 1993; 15 - Georges and Massey 1992; 16 - Kappes, Keele *et al.* 1997; 17 - Kaukinen and Varvio 1993; 18 - Kemp, Hishida *et al.* 1995; 19 - Kossarek, Grosse *et al.* 1993; 20 - Kossarek, Finlay *et al.* 1994; 21 - Kossarek, Grosse *et al.* 1994; 22 - McGraw, Grosse *et al.* 1997; 23 - Mezzelani, Zhang *et al.* 1995; 24 - Moore, Barendse *et al.* 1992; 25 - Moore and Byrne 1993; 26 - Moore, Byrne *et al.* 1994; 27 - Morkos, Grosz *et al.* 1994; 28 - Skow, Goy *et al.* 1994; 29 - Smith, Lopez-Corrales *et al.* 1997; 30 - Steffen, Eggen *et al.* 1993; 31 - Stone, Pulido *et al.* 1995; 32 - Sun, Hart *et al.* 1993; 33 - Sun, Whallon *et al.* 1993; 34 - Toldo, Fries *et al.* 1993; 35 - Vaiman, Osta *et al.* 1992; 36 - Vaiman, Mercier *et al.* 1994; 37 – Georges *et al.* 1993

Population-Locus combinations typed and used for inter-locus comparisons

			Indian zebu				European			W African		
Locus	Classification	Chromosome	Sahiwal	Hariana	Nellore	Ongole	Swiss H-F	Irish H-F	Charolais	N'Dama	Benin Somba	Togo Somba
BMS468	MHC	23			1	1		1	1	1	1	
BOLA-DRB1	MHC	23			1	1		1	1	1	1	
CYP21	MHC	23			1	1		1	1	1	1	
DRB3	MHC	23			1	1		1	1	1	1	
TAMLS113.3	MHC	23			1	1		1	1	1	1	
HBB	Gene	15	2	2				2	2	2	5	
HMH1R	Gene	22	2	2				2	2	2	5	
IL4	Gene	7			1	1		1	1	1		1
OCAM	Gene	29	2	2				2	2	2	5	
PRL	Gene	23	2	2				2	2	2	5	
RASA	Gene	7	2	2				2	2	2	5	
RBP3	Gene	28	2	2				2	2	2	5	
TGLA116	Gene	4	2	2				2	2	2	5	
BL1001	Anonymous	2			1	1		1	1	1		1
BL1067	Anonymous	7			1	1		1	1	1		1
BM1258	Anonymous	23			1	1		1	1	1	1	
BM1815	Anonymous	23			1	1		1	1	1	1	
BM1818	Anonymous	23			1	1		1	1	1	1	
BM1824	Anonymous	1			3	3	4		3	3		
BM2113	Anonymous	2	2	2				2	2	2	5	
BM3010	Anonymous	2			1	1		1	1	1		1
BM6105	Anonymous	7			1	1		1	1	1		1
BM7233	Anonymous	23			1	1		1	1	1	1	
BM733	Anonymous	5			1	1		1	1	1		1
BM8126	Anonymous	5			1	1		1	1	1		1
BMS1116	Anonymous	7			1	1		1	1	1		1
BMS1300	Anonymous	2			1	1		1	1	1		1
BMS2024	Anonymous	2			1	1		1	1	1		1
BMS2269	Anonymous	23			1	1		1	1	1	1	
BMS597	Anonymous	5			1	1		1	1	1		1
BMS713	Anonymous	7			1	1		1	1	1		1
BMS772	Anonymous	5			1	1		1	1	1		1
BMS803	Anonymous	2			1	1		1	1	1		1
BP34	Anonymous	23			1	1		1	1	1	1	

Locus	Classification	Chromosome	Indian zebu				European			W African		
			Sahiwal	Hariana	Nellore	Ongole	Swiss H-F	Irish H-F	Charolais	N'Dama	Benin Somba	Togo Somba
CSRM60	Anonymous	10			3	3	4		3	3		
CSSM005	Anonymous	23			1	1		1	1	1	1	
CSSM024	Anonymous	23			1	1		1	1	1	1	
CSSM042	Anonymous	2			1	1		1	1	1		1
CSSM066	Anonymous	14			3	3	4		3	3		
ETH10	Anonymous	5			1	1	4		1	1		1
ETH131	Anonymous	21	2	2				2	2	2	5	
ETH152	Anonymous	5	2	2				2	2	2	5	
ETH2	Anonymous	5			1	1		1	1	1		1
ETH225	Anonymous	9			1	1		1	1	1		1
HEL1	Anonymous	15			1	1		1	1	1		1
Hel13	Anonymous	11			3	3	4		3	3		
HEL5	Anonymous	21	2	2				2	2	2	5	
HEL9	Anonymous	8			1	1	1		1	1		1
ILSTS001	Anonymous	7			1	1		1	1	1		1
ILSTS005	Anonymous	10			1	1		1	1	1		1
ILSTS006	Anonymous	7			3	3	4		3	3		
ILSTS014	Anonymous	19	2	2				2	2	2	5	
ILSTS030	Anonymous	2			1	1		1	1	1		1
INRA005	Anonymous	12			1	1	4		1	1		1
INRA023	Anonymous	3			3	3	4		3	3		
INRA032	Anonymous	11			3	3		2	2	2	5	
INRA063	Anonymous	18			1	1	4		1	1		1
MGTG7	Anonymous	23			1	1		1	1	1	1	
RM006	Anonymous	7			1	1		1	1	1		1
RM012	Anonymous	7			1	1		1	1	1		1
RM033	Anonymous	23			1	1		1	1	1	1	
RM185	Anonymous	23			1	1		1	1	1	1	
RM356	Anonymous	2			1	1		1	1	1		1
TGLA303	Anonymous	7			1	1		1	1	1		1
TGLA377	Anonymous	2			1	1		1	1	1		1
TGLA48	Anonymous	7			1	1		1	1	1		1
UWCA1	Anonymous	23			1	1		1	1	1	1	

Table 2.3: Population-Locus combinations typed and used for inter-locus comparisons

'H-F' in 'Swiss H-F' and 'Irish H-F' is short for Holstein-Friesian.

Shaded boxes indicate combinations typed. Numbers (1 to 5) indicate source of data:

1 - typed in this study, 2 - data from MacHugh *et al.* (1997), 3 - data from Loftus *et al.* (1999), 4 - data from Schmid *et al.* (1999), 5 - C. Meghen, unpublished data

2.3.3 Hardy-Weinberg equilibrium

Of the 396 population-locus combinations in the dataset (table 1), 392 are polymorphic. These were tested for deviation from Hardy-Weinberg equilibrium proportions using a version of Fisher's exact test.

	Population-locus combinations departing significantly from HWE	
Probability (*p*)	**Observed**	**Expected by chance**
0.001	1	0.4
0.01	6	4
0.05	27	19

Table 2.4: Departure from Hardy-Weinberg equilibrium for all population-locus combinations in dataset assessed using an exact test

The results (**Table 2.4**) indicate that the number of population-locus combinations deviating from Hardy-Weinberg equilibrium is slightly higher than expected. However, none of the individual probability values is significant when confidence intervals are corrected for multiple testing using the sequential form of the Bonferroni correction (Rice 1989). Probabilities for individual population-locus combinations can be combined using Fisher's method of combining probabilities. According to this method, if r independent tests are performed, an overall Chi square value, χ^2_{TOT}, distributed with 2 x r degrees of freedom, can be calculated according to the formula:

$$\chi^2_{TOT} = -2\sum_{j=1}^{r} \ln P_j$$

Equation 2.10

where P_j is the probability value for the j'th test. Here, a total Chi square value of 827 with 784 d.f. gives an overall probability of 0.14, suggesting that there is no significant overall departure from HWE.

Population	Population-locus combinations typed	Population-locus combinations giving HWE exact test probabilities below p = 0.05	
		Observed	Expected by chance
Sahiwal	12	0	0.6
Hariana	12	0	0.6
Nellore	55	3	2.75
Ongole	55	7	2.75
Holstein	11	0	0.55
Holstein-Friesian	56	4	2.8
Charolais	66	5	3.3
N'Dama	65	6	3.25
Benin Somba	28	1	1.4
Togo Somba	32	1	1.6

Table 2.5: Number of loci showing departure from HWE proportions in individual populations

Considering each population separately (**Table 2.5**), the greatest proportion of population-locus combinations showing departure from HWE is seen in Ongole. However, no populations show significant departure from HWE when test probabilities are combined across loci using Fisher's method.

Locus	Indian zebu				European			W African		
	Sahiwal	Hariana	Nellore	Ongole	Swiss H-F	Irish H-F	Charolais	N'Dama	Benin Somba	Togo Somba
BOLA-DRB1	2	2	1	1		1	1	1	1	
BM1818			1	1	4	1	1	1	1	
BM2113	2	2	3	3	4	2	2	2	5	
CSSM024			1	1		1	1	1	1	1
ETH152	2	2	3	3	4	2	2	2	5	
ETH225	2	2	1	1	4	1	1	1	1	1
HEL1	2	2	1	1	4	1	1	1	1	1
HEL5	2	2	3	3	4	2	2	2	5	
ILSTS001	2	2	1	1		1	1	1	5	1
ILSTS005	2	2	1	1	4	1	1	1	5	1
INRA032	2	2	3	3	4	2	2	2	5	
TGLA48	2	2	1	1		1	1	1	5	1

Table 2.6: Population-locus combinations typed and used for comparison of 'equivalent' populations

'H-F' in 'Swiss H-F' and 'Irish H-F' is short for Holstein-Friesian.

Shaded boxes indicate combinations typed. Numbers (1 to 5) indicate source of data:

1 - typed in this study, 2 - data from MacHugh *et al.* (1997), 3 - data from Loftus *et al.* (1999), 4 - data from Schmid *et al.* (1999), 5 - C. Meghen, unpublished data

2.3.4 Use of 'equivalent' populations in creating dataset for analysis

Ideally, for inter-locus comparisons, the same samples should be typed for each locus. This is not the case here, as many loci were not typed in specific populations. However, all loci were typed in at least two European taurine breeds and at least two Indian zebu breeds whilst for West African taurine cattle, 61 of the 67 loci were typed in two breeds with the remaining six loci only typed in one breed. This means that in almost all cases where a locus was not typed in a particular population, it was typed in a similar population. Determining the genetic distance between these similar populations using loci typed in both reveals whether they can be considered as equivalent.

The Irish Holstein-Friesian (H-F) population was not typed for ten of the 67 loci in the dataset, although all ten loci have been typed in a Swiss H-F population. To determine

whether the Swiss H-F data could be substituted for the missing Irish H-F data, the populations were compared for seven loci typed in both (**Table 2.6**). For the seven loci, Nei's standard genetic distance between the two H-F populations is 0.018 ± 0.01, approximately ten-fold lower than the corresponding Swiss H-F-Charolais (0.21 ± 0.09) or Irish H-F-Charolais (0.20 ± 0.09) values.

Similarly, about half of the loci were typed in Benin Somba, and half in Togolese Somba. Six loci were typed in both populations (**Table 2.6**), and for these loci, the Togo-Benin Somba standard genetic distance is $3x10^{-4}$ ± $5x10^{-3}$, substantially lower than the corresponding Benin Somba-N'Dama (0.071 ± 0.05) or the Togo Somba-N'Dama (0.084 ± 005) values.

With regard to the zebu populations, 12 loci were not typed in Ongole or Nellore, but were typed in Hariana and Sahiwal. Ten loci were typed in all four populations, and also in several *B. taurus* populations (**Table 2.6**). Pairwise standard genetic distance values are shown in **Table 2.7**. Ongole and Nellore are genetically more similar to one another than to either Hariana or Sahiwal, and vice versa. However, all of the zebu populations are genetically much closer to one another than to European or African *B. taurus* populations (**Table 2.7**).

	Hariana	Nellore	Ongole	Sahiwal	Holstein-Friesian
Nellore	0.11 ± 0.05				
Ongole	0.13 ± 0.06	0.07 ± 0.02			
Sahiwal	0.06 ± 0.04	0.13 ± 0.06	0.17 ± 0.08		
Holstein-Friesian	1.44 ± 0.32	1.70 ± 0.31	1.55 ± 0.38	1.11 ± 0.32	
N'Dama	1.37 ± 0.30	1.77 ± 0.40	1.41 ± 0.30	1.35 ± 0.38	0.47 ± 0.18

Table 2.7: Genetic relationships between 4 *Bos indicus* and 2 *Bos taurus* populations (Nei's standard genetic distance) typed for ten microsatellite loci

2.3.5 Approach to analysis in view of findings on population equivalence

The genetic distance results suggest that there is very little genetic differentiation between the Irish and Swiss H-F populations implying that data from the Swiss population can be used to substitute for missing Irish data. Similarly, the very close genetic relationship between the Togolese and Benin Somba populations means that data from one can be used to substitute for missing data in the other.

It is less clear, however, that the Hariana and Sahiwal can be considered as equivalent to the Ongole and Nellore populations. Therefore caution should be exercised when making inter-locus comparisons using data for loci typed in Ongole and Nellore and loci typed in Hariana and Sahiwal. For this reason, all subsequent inter-locus analyses were performed using the entire dataset - i.e. data for all population-locus combinations listed in **Table 2.3** - and also using two sub-datasets. The first of these included only data for the taurine breeds for the full 67 loci, and will be referred to as the taurine sub-dataset. The second sub-dataset excluded all 12 loci typed in Hariana and Sahiwal. As these 12 loci included seven of the eight gene-linked loci, the final gene-linked locus (IL4) was also excluded, leaving 54 loci (49 anonymous and 5 MHC-linked) typed in the taurine breeds and Ongole and Nellore. This second sub-dataset will be referred to as the 54-locus sub-dataset.

2.3.6 Intra-population genetic diversity

To determine whether genomic background influences intra-population microsatellite diversity, loci were compared for two measures of genetic diversity; the number of distinct alleles per locus, and Nei's gene diversity (expected heterozygosity).

2.3.7 Number of distinct alleles per locus

The number of distinct alleles per locus is highly sensitive to sample size (Nei 1987). Larger sample sizes tend to give higher values. A resampling strategy was used to illustrate this relationship. For the two breeds typed for all 67 loci (Charolais and N'Dama) a given number of alleles were resampled with replacement at each locus, and the number of distinct alleles in the sample was determined. 100 replicate resamplings were performed, and the results were averaged across replicates and loci. The number of alleles sampled was varied between 10 and 50. Results are shown in **Figure 2.4**. For both breeds, the mean number of alleles per locus increases sharply with allele size when the number of alleles sampled is small. The rate of increase declines as the number of alleles increases, but there remains a positive association even when the number of alleles sampled is as high as 50.

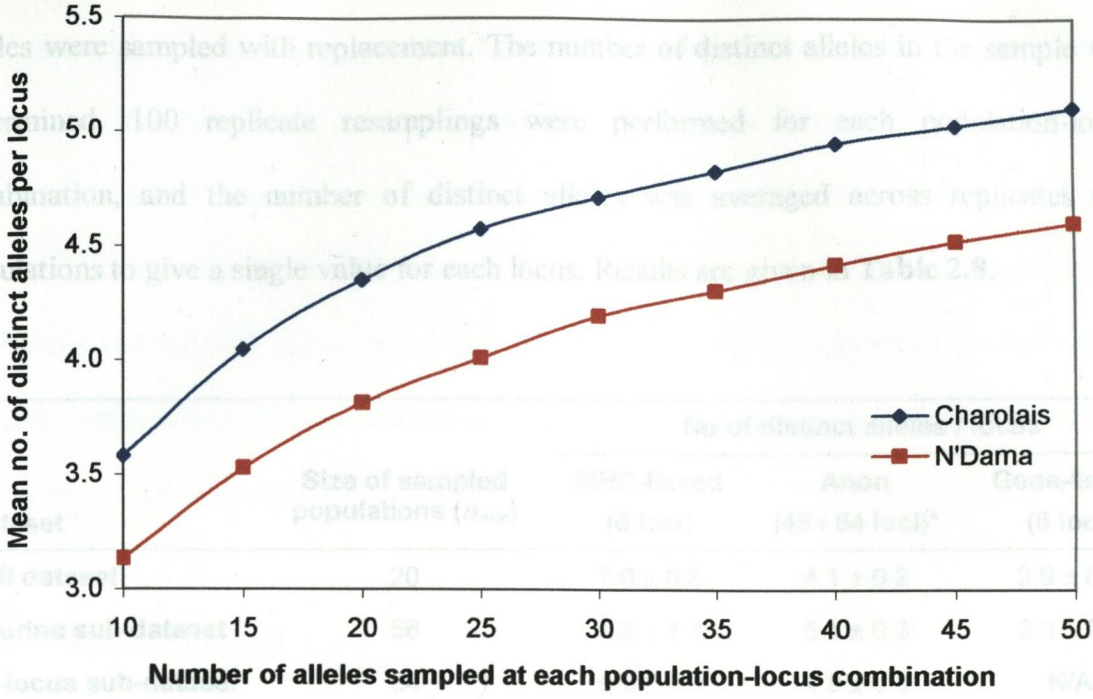

Figure 2.4: Dependence of number of distinct alleles per locus on sample size
Each data point is the average across 100 replicate resamplings of alleles from either Charolais or N'Dama at 67 microsatellite loci.

In the dataset assembled here, there is variation in the sample sizes of the different populations used, with sample sizes for Hariana and Sahiwal considerably lower than for the other populations (**Table 2.1**). In addition to this inter-population sample-size variation there is also intra-population variation in the number of samples successfully amplified at each locus.

To correct for the effects of sample size variation, the smallest number of samples successfully amplified at any population-locus combination, n_{min}, was determined. For the full dataset, n_{min} is 20 (for the loci typed in Hariana). For the 54-locus sub-dataset (excluding loci typed in Hariana and Sahiwal), n_{min} is 54. For the taurine sub-dataset (excluding all zebu breeds), n_{min} is 56.

For every population-locus combination in the full dataset and the two sub-datasets, n_{min} alleles were sampled with replacement. The number of distinct alleles in the sample was determined. 100 replicate resamplings were performed for each population-locus combination, and the number of distinct alleles was averaged across replicates and populations to give a single value for each locus. Results are given in **Table 2.8.**

		No of distinct alleles / locus		
Dataset	**Size of sampled populations (n_{min})**	**MHC-linked (5 loci)**	**Anon (49 / 54 loci)***	**Gene-linked (8 loci)**
Full dataset	20	7.0 ± 0.8	4.1 ± 0.2	2.9 ± 0.4
Taurine sub-dataset	56	9.3 ± 1.4	5.1 ± 0.3	3.3 ± 0.5
54-locus sub-dataset	54	9.0 ± 1.3	4.9 ± 0.3	N/A

Table 2.8: Effects of genomic background on the number of distinct alleles at microsatellite loci

No of distinct alleles was estimated by performing 100 samplings of n_{min} alleles at each population-locus combination, and averaging the number of distinct alleles across replicates, populations and loci. *The figure for anonymous loci is the average over 54 loci for the full dataset and the taurine sub-dataset and the average over 49 loci for the 54-locus sub-dataset.

The observed differences between the three classes of microsatellite in the number of distinct alleles per locus were tested for significance. Taking results for the full dataset, a one-way ANOVA with three classes of loci (MHC-linked, gene-linked and anonymous) showed a highly significant difference ($p<0.001$) between the locus classes. Post-hoc tests (Tukey's HSD) show that the MHC-linked loci have significantly higher numbers of distinct alleles per locus than either the gene-linked loci ($p<0.001$) or the anonymous loci ($p<0.001$). Additionally, the gene-linked loci have significantly fewer alleles per locus than the anonymous loci ($p=0.02$). The analogous non-parametric (Kruskal-Wallis) test also indicates highly significant ($p<0.001$) differences between locus classes. Similar significance results are observed when the two sub-datasets are analysed. For the taurine sub-dataset, a one-way ANOVA again shows a highly significant difference between the

three classes of loci ($p<0.001$), with post-hoc (Tukey's HSD) tests confirming that the MHC-linked loci have higher numbers of alleles than either the gene-linked ($p<0.001$) or anonymous loci ($p<0.001$). Again, the number of distinct alleles at gene-linked loci is significantly lower than at anonymous loci ($p=0.04$). The non-parametric Kruskal-Wallis test also indicates significant differences between locus classes ($p<0.001$). When loci typed in Hariana and Sahiwal are excluded (54-locus sub-dataset), the number of distinct alleles at MHC loci remains significantly higher than at anonymous loci (t test assuming unequal variances $p=0.04$, Mann-Whitney U test $p=0.001$).

2.3.8 Gene Diversity

Gene diversity was calculated for each population-locus combination in the full dataset, and in the two sub-datasets. Results were averaged across populations to give a single value for each locus. Average gene diversities for each class of microsatellite locus are given in **Table 2.9**.

	Gene diversity		
Dataset	**MHC-linked (5 loci)**	**Anon (49 / 54 loci)***	**Gene-linked (8 loci)**
Full dataset	0.83 ± 0.02	0.62 ± 0.02	0.52 ± 0.08
Taurine sub-dataset	0.84 ± 0.02	0.60 ± 0.02	0.42 ± 0.08
54-locus sub-dataset	0.83 ± 0.02	0.62 ± 0.03	N/A

Table 2.9: Effects of genomic background on gene diversity (expected heterozygosity) at microsatellite loci

Gene diversity was calculated for each population-locus combination, and averaged across populations and loci. *The figure for anonymous loci is the average over 54 loci for the full dataset and the taurine sub-dataset and the average over 49 loci for the 54-locus sub-dataset.

The observed differences in gene diversity between the three classes of microsatellite were tested for significance. Using the full dataset, a one-way ANOVA showed a highly

significant difference ($p<0.001$) between the locus classes. Post-hoc tests (Tukey's HSD) show that the gene diversity at MHC-linked loci is significantly higher than at either the gene-linked loci ($p<0.001$) or the anonymous loci ($p=0.001$). Gene diversity at gene-linked loci is significantly lower than at anonymous loci ($p=0.04$). The analogous non-parametric (Kruskal-Wallis) test also indicates highly significant ($p<0.001$) differences between locus classes. Similar significance results are observed when the two sub-datasets are analysed. For the taurine sub-dataset, a one-way ANOVA again shows a highly significant difference between the three classes of loci ($p<0.001$), with post-hoc (Tukey's HSD) tests confirming significantly higher gene diversity at MHC-linked loci than at gene-linked ($p<0.001$) or anonymous loci ($p=0.002$), and significantly higher gene diversity at anonymous loci than at gene-linked loci ($p=0.003$). The non-parametric Kruskal-Wallis test also indicates significant differences in gene diversity between locus classes ($p<0.001$). When loci typed in Hariana and Sahiwal are excluded (54-locus sub-dataset), gene diversity at MHC loci remains significantly higher than at anonymous loci (t test assuming unequal variances $p<0.001$, Mann-Whitney U test $p<0.001$).

2.3.9 Comparison of observed and neutral allele distributions

The two measures of genetic diversity described (number of distinct alleles per locus and gene diversity) are related. Using data for all population-locus combinations, there is a very highly significant positive correlation between the number of distinct alleles and heterozygosity (Pearson r coefficient 0.60, 394 d.f., $p<0.001$). When the MHC and gene-linked loci are excluded, the correlation remains (Pearson r coefficient 0.66, 316 d.f., $p<0.001$). It is therefore unsurprising that differences between classes of microsatellite revealed using one statistic should also be observed using the other. It is, however, possible to predict the level of gene diversity for a locus at MDE with a given number of distinct alleles (Ewens 1972). Comparing the predicted gene diversity value with the observed

value should then give a measure of departure from MDE that is independent of the number of distinct alleles.

There are a number of possible causes of departure from mutation-drift equilibrium. These include population admixture (Chakraborty 1990), population bottlenecks or rapid expansion (Cornuet and Luikart 1996) and selection, whether heterozygote advantage or disadvantage or disadvantage due to deleterious alleles (Watterson 1977). Selection, our primary concern here, should act independently at independent loci, whereas admixture and bottlenecks or population expansion should affect all loci. Watterson has shown that the statistic F_{obs}, equal to the sum of squared allele frequencies (or the expected homozygosity) for the sample, can be compared with the predicted sum of squared allele frequencies under MDE, F_{exp}, to indicate whether a locus is under selection (Watterson 1977). In the case of heterozygote advantage, where all heterozygotes are selectively favoured over homozygotes, F tends to be depressed, and allele frequencies relatively even. In contrast, homozygote advantage leads to elevated F and more variable allele frequencies (Watterson 1977). Selection against deleterious alleles is also predicted to increase F and to increase variability in allele frequencies. However, the influence of heterozygote advantage or disadvantage is predicted to be much greater than that of deleterious-allele selection (Watterson 1978).

Two methods were used to estimate expected F_{exp} as described in materials and methods. The first employed a strict stepwise mutation model (SMM), and the second the infinite allele mutation model (IAM). When the SMM is used to estimate F_{exp}, there is a positive correlation between the number of distinct alleles and F_{obs}/F_{exp}. This correlation is seen both for data from all loci (Pearson r coefficient 0.289, 391 d.f., $p<0.001$) and when MHC-linked and gene-linked loci are excluded (Pearson r coefficient 0.419, 314 d.f., $p<0.001$).

Given that there are significant differences in the number of distinct alleles between the three classes of microsatellite studied here, inter-class comparison based on estimates of F_{exp} obtained under the SMM is inappropriate. In contrast, when estimates of F_{exp} are obtained under the IAM, there is no correlation between the number of distinct alleles and F_{obs}/F_{exp} (Pearson r coefficient -0.047 with 391 d.f., p=0.35 for all loci; Pearson r coefficient 0.098 with 314 d.f., p=0.08 for 'anonymous' loci only). Hence the infinite allele model estimates of F_{exp} were used in all subsequent analysis.

Watterson has proposed a test of selective neutrality based on the statistic F (Watterson 1977). The test statistic proposed by Watterson is the cumulative probability of allele distributions, drawn from the Ewens' neutral distribution, whose F value (F_{exp}) is less than or equal to the observed F value (F_{obs}) (Watterson 1977). Low probabilities indicate heterozygote advantage, whilst high probabilities indicate homozygote advantage or deleterious-allele selection. However, a problem with the use of the probability test in this is that the infinite allele mutation model assumed is not wholly appropriate for microsatellite data. Using IAM estimates of F_{exp}, the average of F_{obs}/F_{exp} across all population-locus combinations is 0.82, implying an excess of gene diversity across all loci. When the SMM is used to estimate F_{exp}, average F_{obs}/F_{exp} is 1.2, suggesting a deficit of gene diversity. Clearly the value of F_{exp} is highly dependent on the model assumed in estimating it, hence it would be wrong to take literally the significance values returned by the Watterson test. A Kruskal-Wallis rank test does, however, show that Watterson test probabilities for gene-linked, MHC-linked and anonymous loci are significantly different ($p<0.001$). Instead of using test probabilities, we have chosen to report values of F_{obs}/F_{exp} for each locus. While no significance is attached to the absolute value of F_{obs}/F_{exp} the large panel of loci assembled here permits inter-locus comparisons of F_{obs}/F_{exp} to determine whether selection is acting at particular loci.

Dataset	F_{obs} / F_{exp} (IAM assumed)		
	MHC-linked (5 loci)	Anon (49 / 54 loci)*	Gene-linked (8 loci)
Full dataset	0.62 ± 0.03	0.82 ± 0.02	0.91 ± 0.04
Taurine sub-dataset	0.58 ± 0.03	0.83 ± 0.02	0.92 ± 0.07
54-locus sub-dataset	0.62 ± 0.03	0.82 ± 0.02	N/A

Table 2.10: Effects of genomic background on departure from MDE at microsatellite loci
Homozygosity was calculated for each population-locus combination, and divided by the estimated expected homozygosity under the IAM. Results were averaged across populations and loci. *The figure for anonymous loci is the average over 54 loci for the full dataset and the taurine sub-dataset and the average over 49 loci for the 54-locus sub-dataset.

To investigate associations between locus background and excess or deficiency of gene diversity, F_{exp} was estimated using the IAM for each population-locus combination for each of the three classes of microsatellite studied, and F_{obs}/F_{exp} was calculated and averaged across populations and loci for each class. Results, shown in **Table 2.10**, indicate that, compared with anonymous loci, MHC-linked loci show an excess of gene diversity whereas gene-linked loci show a deficiency of gene diversity. The nature of heterozygosity excess and deficiency are illustrated in **Figure 2.5**, which shows allele frequency distributions for an MHC-linked microsatellite (BMS468) and a gene-linked microsatellite (RBP3) typed in Charolais and Holstein-Friesian cattle. Although the number of distinct alleles is comparable (5 for BMS468 against 4 for RBP3), the shapes of the frequency distributions for the two loci are very different. At BMS468, allele frequencies are close to equal, whereas at RBP3, there is one modal allele with a frequency more than ten times that of the next most common allele.

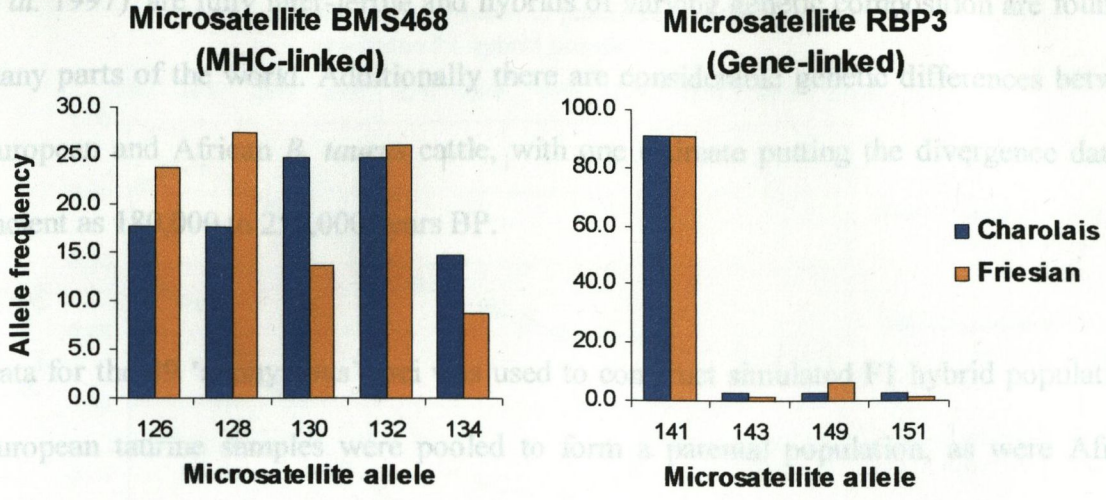

Figure 2.5: Allele frequency distributions for MHC-linked and gene-linked microsatellites typed in two European cattle breeds

The difference in F_{obs}/F_{exp} between MHC-linked gene-linked and anonymous loci (**Table 2.10**) is highly significant, whether the full dataset (ANOVA $p<0.001$, Kruskal-Wallis $p<0.001$) or the taurine sub-dataset is used (ANOVA $p=0.001$, Kruskal-Wallis $p=0.001$). Further analysis shows that F_{obs}/F_{exp} is significantly lower at MHC-loci than at anonymous loci ($p=0.001$ for both datasets, Tukey's HSD test). The difference between anonymous and gene-linked loci, though, is not significant ($p=0.137$ for all data, $p=0.284$ for taurine data). Similarly, using the 49-locus sub-dataset, F_{obs}/F_{exp} is significantly lower for MHC-linked loci than for anonymous loci (t test $p=0.001$, Mann-Whitney U test $p<0.001$).

2.3.10 Effects of population admixture on allele frequency distributions

To further demonstrate the need for caution in inferring selection based on Watterson's F statistic, simulations were undertaken to investigate the effects of population admixture on F. It has been noted that admixture between genetically distinct populations can perturb mutation-drift equilibrium in the resulting hybrid population (Chakraborty 1990). This situation is particularly relevant in cattle. The two sub-species *B. taurus* and *B. indicus*, which diverged an estimated 200,000 to 1,000,000 years BP (Loftus *et al.* 1994; MacHugh

et al. 1997), are fully inter-fertile and hybrids of varying genetic composition are found in many parts of the world. Additionally there are considerable genetic differences between European and African *B. taurus* cattle, with one estimate putting the divergence date as ancient as 180,000 to 250,000 years BP.

Data for the 49 'anonymous' loci was used to construct simulated F1 hybrid populations. European taurine samples were pooled to form a parental population, as were African taurine samples and Indian zebu samples. Two parental populations were selected at a time, and simulated F1 hybrid populations of 80 alleles were created by sampling alleles from the parental populations. The percentage of admixture was varied in increments from 100% parental population 1 (0% population 2) to 0% parental population 1 (100% parental population 2). 50 replicate populations were created for each degree of admixture. F_{obs}/F_{exp} was calculated for each population, with F_{exp} estimated under the IAM. Results are shown in **Figure 2.6**.

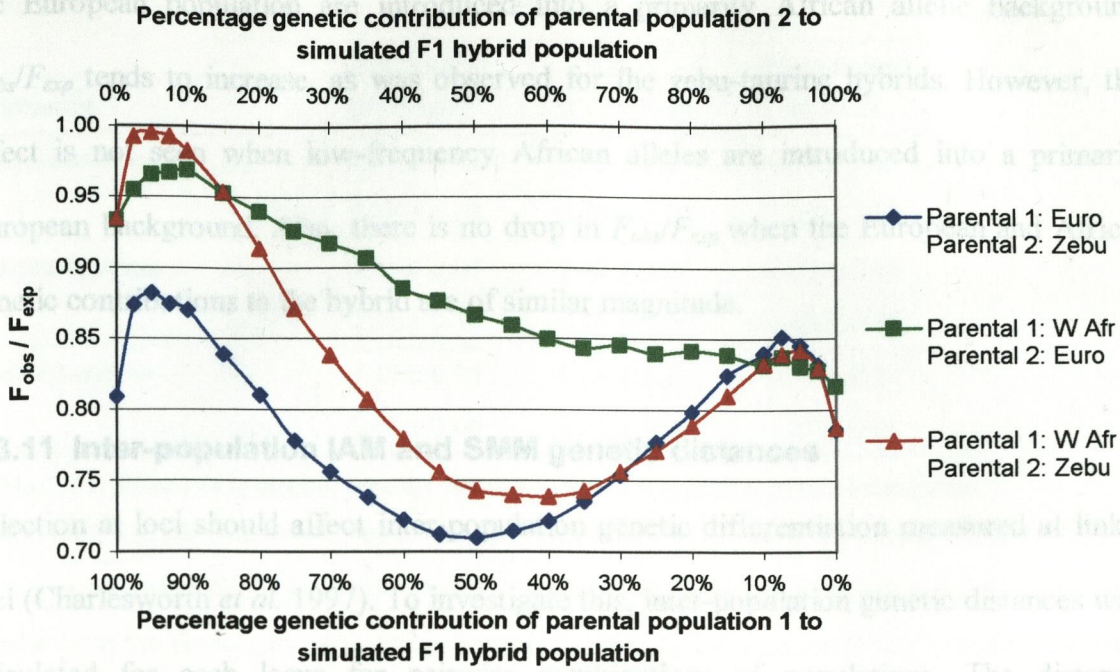

Figure 2.6: Effects of population admixture on departure from MDE at microsatellites
Each data point is based on 50 replicate simulated F1 hybrid populations of 40 individuals typed at 49 microsatellite loci, constructed by sampling alleles from parental populations with replacement. The three parental populations used are the pooled European taurine breeds (Euro), pooled West African breeds (W Afr) and the pooled Indian zebu breeds (Zebu).

When simulated hybrid populations were constructed using the zebu population and either of the taurine populations as parental populations, a low level of admixture (<20%) leads to an increase in F_{obs}/F_{exp} above the level seen in the principal contributing parental population. This is due to the introduction of novel alleles from the second parental population. These alleles are at low frequency in the hybrid, resulting in an allele distribution for the hybrid that is more uneven that that of the principal parental population. However, when the hybrid population has comparatively equal genetic contributions from the two parental populations, F_{obs}/F_{exp} tends to decrease below the level seen in the parental populations. This is because the combination of two populations with quite distinct allele distributions results in a hybrid that has a more even allele distribution. The effects of hybridisation are different when the hybrid populations are constructed from the more closely related European and African populations. When a small number of alleles from

the European population are introduced into a primarily African allelic background, F_{obs}/F_{exp} tends to increase, as was observed for the zebu-taurine hybrids. However, this effect is not seen when low-frequency African alleles are introduced into a primarily European background. Also, there is no drop in F_{obs}/F_{exp} when the European and African genetic contributions to the hybrid are of similar magnitude.

2.3.11 Inter-population IAM and SMM genetic distances

Selection at loci should affect inter-population genetic differentiation measured at linked loci (Charlesworth *et al.* 1997). To investigate this, inter-population genetic distances were calculated for each locus for pairwise combinations of populations. The distances calculated were Nei's standard genetic distance D_S (Nei 1972), derived under the infinite allele model (IAM), and Goldstein *et al.*'s delta mu squared distance $(\delta\mu)^2$ (Goldstein *et al.* 1995b), derived under the stepwise mutation model (SMM). In calculating the latter distance, only loci with repeat spectra consistent with the SMM were considered. Eight loci were rejected due to alleles that were separated by non-integer numbers of repeats. Six of the 54 anonymous loci were rejected, and two of the five MHC-linked loci.

		Genetic distance		
Distance		**MHC-linked (5 / 3 loci)***	**Anon (54 / 48 loci)***	**Gene-linked (8 loci)**
Intra-continental	D_S	0.29 ± 0.07	0.10 ± 0.01	0.04 ± 0.03
	$(\delta\mu)^2$	0.22 ± 0.26	0.48 ± 0.11	0.08 ± 0.03
Africa-Europe	D_S	0.24 ± 0.07	0.49 ± 0.07	0.22 ± 0.13
	$(\delta\mu)^2$	0.25 ± 0.16	1.45 ± 0.25	0.59 ± 0.36
Zebu-Taurine	D_S	0.43 ± 0.11	1.73 ± 0.23	0.86 ± 0.22
	$(\delta\mu)^2$	0.42 ± 0.05	6.00 ± 1.27	1.62 ± 0.76

Table 2.11: Effect of genomic background on inter-population genetic distance

Two different genetic distance measures were estimated: Nei's standard genetic distance (D_S) and delta mu squared $(\delta\mu)^2$. Intra-continental distance is the average of pairwise distances between breeds from the same continent. Africa-Europe distance is the average of four pairwise distances between African and European breeds. Zebu-Taurine distance is the average of eight pairwise distances between zebu and taurine breeds. *Loci with alleles separated by non-integer numbers of repeats were excluded in calculating $(\delta\mu)^2$.

Nei's D_S is expected to be linear over time for loci evolving under the IAM, and not for loci evolving under the SMM. Conversely, Goldstein *et al.*'s $(\delta\mu)^2$ is expected to be linear over time for loci evolving under the SMM, and not the IAM. The mode of evolution for microsatellites is believed to closely resemble the SMM, hence it might appear that $(\delta\mu)^2$ is the more appropriate statistic for measuring genetic distance. However, inspection of the microsatellite sequences (**Table 2.2**) shows that approximately half do not conform to the expected perfect repeat pattern for loci evolving under the SMM. Instead they have complex repeats consisting of multiple arrays of different repeat units or interrupted repeats. It is possible that the mutation process at such loci deviates from the SMM. Also occasional multi-step mutations (Di Rienzo *et al.* 1994) are known to occur. Any significant departure from the SMM would improve the performance of classical genetic distance measures such as D_S relative to $(\delta\mu)^2$ (Goldstein *et al.* 1995a; Slatkin 1995b). A separate point is that, for populations that have recently diverged, such as the populations

from within the same continent sampled here, the role of genetic drift in evolution would assume greater importance relative to mutation, again improving the performance of D_S (Slatkin 1995b).

The greatest genetic distances are those between zebu and taurine populations. Distances between African and European taurine breeds are lower, and distances between breeds from the same continent lower still. These results agree with our understanding of the cattle phylogeny. It is noticeable though that both the IAM and SMM distances show comparable increases as we move from the most closely to the most distantly-related populations. This suggests that the mode of evolution for the loci studied is intermediate between the IAM and SMM models.

Genetic distances calculated using gene-linked loci are consistently lower than distances calculated using anonymous loci (**Table 2.11**). This applies for both IAM and stepwise distances between breeds from the same continent, between African and European breeds, and between zebu and taurine breeds. Distances calculated using MHC-linked loci are also lower than distances calculated using anonymous loci, but with a single exception; the average D_S distance between breeds from the same continent is higher for MHC-linked loci (**Table 2.11**). Comparing distances calculated using gene-linked loci with those calculated using MHC-linked loci, we find that estimates of distances between the distantly related zebu and taurine breeds are greater when calculated using gene-linked loci. However, the situation is reversed for distance estimates for closely related breeds from the same continent.

Differences in genetic distance between classes of microsatellite loci were tested for significance. It should be noted that there were only three $(\delta\mu)^2$ values for MHC loci,

hence there is limited potential for detection of significant differences between MHC-linked and other loci.

When D_S distances between breeds from the same continent are compared, there is a highly significant difference between locus classes (one-way ANOVA $p < 0.001$, Kruskal-Wallis $p = 0.001$). This difference arises because the distance for MHC-linked loci is significantly lower than for either gene-linked loci or anonymous loci (Tukey's HSD test, $p<0.001$). Intra-continental D_S values for MHC-linked loci remain higher than for other loci when the zebu populations are excluded (due to possible non-equivalence between zebu populations typed at different loci, as discussed). In contrast, no significant difference between locus classes is seen for the average SMM distance $(\delta\mu)^2$ between breeds from the same continent. When distances between African and European populations are compared, there is no significant difference between locus classes for either the D_S or $(\delta\mu)^2$ distances. Finally, comparing Nei's D_S values for average zebu-taurine distances, a one-way ANOVA gives a non-significant difference between microsatellite locus classes ($p=0.09$), whereas the non-parametric Kruskal-Wallis test indicates a significant difference ($p=0.018$). Testing the corresponding delta-mu squared distances reveals no significant differences between locus classes.

2.4 DISCUSSION

This study aimed to investigate the viability of detecting genes under selection through effects on mapped linked microsatellite loci. We have tested predictions concerning the effects of different modes of selection acting at genes on linked neutral loci and found evidence to suggest that variability at bovine microsatellites may indeed be influenced by selection.

2.4.1 Evidence of selection at MHC-linked microsatellites

As discussed earlier, theory suggests that microsatellites in regions of balancing selection may show increased allelic diversity, unusually even allele frequency distributions and reduced genetic differentiation between distinct populations. All of these predicted effects are observed in microsatellites in the bovine MHC region.

2.4.2 Increased allelic diversity for MHC-linked microsatellites

The bovine MHC-linked microsatellites studied show higher allelic diversity than microsatellites outside the MHC region. The number of distinct alleles per locus (**Table 2.8**) and the expected heterozygosity (**Table 2.9**) are both higher for MHC-linked microsatellites. Other studies also indicate that MHC-linked microsatellites may exhibit high allelic diversity. Of five microsatellites on ovine chromosome 20 typed in Soay sheep, the two loci with the highest heterozygosity were in the MHC class II region, with one locus in the MHC class I region and two distant loci showing lower heterozygosity (Paterson 1998). Studies of murine microsatellites also find higher numbers of alleles for microsatellites within the MHC than outside (Meagher and Potts 1997).

The observed increase in diversity at MHC-linked loci agrees with simulation results for loci linked to genes under balancing selection. Using an infinite allele model with multiple nucleotide sites, each of which can have one or two alleles, a quantity termed the 'sum of site heterozygosities' can be calculated according to the formula $SSH = \sum_i^S 2p_i(1-p_i)$, where p_i is the frequency of one of the alleles at the i'th site of S segregating sites (Gillespie 1997). In population simulations, *SSH* values for loci near to genes subject to heterozygote advantage are up to threefold higher than for strictly neutral loci (Gillespie 1997). Allelic diversity at a neutral locus linked to a selected locus can be partitioned into two components: variation of alleles at the neutral locus linked to each separate allele at the selected locus (intra-allelic class diversity) and variation of alleles at the neutral locus linked to different alleles at the selected locus (inter-allelic class diversity) (Charlesworth *et al.* 1997). Theory predicts that the intra-allelic component of diversity should be approximately the same for loci influenced by local balancing selection and loci experiencing no selective influence (Charlesworth *et al.* 1997). The inter-allelic component, however, is expected to increase with local balancing selection (Charlesworth *et al.* 1997; Slatkin 1995a), thus increasing overall allelic diversity. Simulation results agree with these theoretical predictions (Charlesworth *et al.* 1997).

2.4.3 Even allele frequency distributions at MHC-linked microsatellites

We find that MHC microsatellites tend to show more even allele distributions than non-MHC loci as measured by tests of neutrality based on Ewens' theory (**Table 2.10**). This implies that selective forces acting to retain diversity in the MHC region reduce the extent of genetic drift at linked microsatellites, and thereby retain microsatellite alleles.

Other studies of MHC microsatellites in ruminants have also reported unusually even allele distributions, indicative of balancing selection. Two of the microsatellites studied here, DRB3 and BoLA DRB1 were reported to show heterozygosity excess in a range of cattle breeds (van Haeringen *et al.* 1999). Samples of 15 breeds, of which 14 were European and one an African *B. taurus* – *B. indicus* hybrid, were typed at each locus. The Ewens-Watterson test of homozygosity was performed for each breed-locus combination, with expected homozygosity estimated under the IAM. A binomial test shows that expected homozygosity exceeds observed homozygosity for a very significant excess of breed-locus combinations. However, as is found here, the test is inappropriate for microsatellite data as the expected homozygosity for a locus evolving under the IAM tends to be higher than for a locus evolving under the SMM. The inclusion of a taurine-zebu hybrid breed is a further confounding factor, as admixture tends to cause deviation from MDE, which could lead to misinterpretation. A second study of 20 Northern European breeds, typed at 10 microsatellites also finds evidence for heterozygosity excess at two loci, of which one is BoLA DRB1 (Kantanen *et al.* 2000). The second locus, CSSM066, ranks 49th out of the 67 loci typed here for heterozygosity excess as measured by F_{obs}/F_{exp} (data not shown). Finally, the study of five ovine chromosome 20 microsatellites typed in Soay sheep shows heterozygosity excess at two of three MHC-linked microsatellites (Paterson 1998). Two distant microsatellites on the same chromosome showed no heterozygosity excess. These two latter studies also both used the Ewens-Watterson homozygosity test assuming the IAM. However, the finding in both studies that non-MHC loci do not show heterozygosity excess confirms the significance of the results for MHC-linked loci.

2.4.4 Low inter-population genetic distance at MHC-linked microsatellites

In this study, we find that MHC-linked loci give low estimates of inter-population genetic subdivision for distantly related populations, but not for closely related populations (**Table**

2.11). This applies to both the IAM distance D_S (**Equation 2.8**) and the SMM distance $(\delta\mu)^2$ (**Equation 2.9**). Strikingly, very little differentiation is seen between *Bos taurus* and *Bos indicus* populations at MHC loci, despite a separation of the order of hundreds of thousands of years between the lineages.

Similar results were obtained in a study of 20 microsatellite loci typed in 20 European, African and Indian cattle breeds. Genetic sub-division between breeds was estimated using Wright's F_{ST} (Wright 1951), Weir and Cockerham's θ (Weir and Cockerham 1984) and Nei's G_{ST} (Nei 1973). These estimators of genetic partitioning give an indication of average genetic subdivision across all pairwise combinations of populations. For the 20 breeds sampled, the majority of pairwise combinations consist of relatively distantly related populations, in many cases from different continents. Comparing results for all 20 loci reveals that inter-population genetic subdivision is lowest for the MHC-linked microsatellites Bola-DRB1 and BoLA-DRB2 (MacHugh 1996).

The results suggest a reduction of allele loss through genetic drift at MHC-linked microsatellites, leading to retardation of inter-population differentiation over time. This phenomenon is well known in MHC gene coding sequences, which show trans-species polymorphism due to preservation of ancient ancestral sequences in divergent daughter species (Klein *et al.* 1993). Sequences of MHC-linked microsatellites from primates provide evidence that trans-species polymorphism is not restricted to coding sequences, but can also apply to microsatellites alleles. Bergstrom *et al.* (1999) sequenced a microsatellite in the second intron of the MHC class II DRB1 gene in humans, chimpanzees and gorillas and found a number of highly divergent repeat motifs. Three of these divergent motifs were seen in both human and chimpanzee, and two in both human and gorilla (Bergstrom *et al.* 1999). This implies that multiple microsatellite motifs were present in the common

ancestral species, and that microsatellite polymorphism has been maintained in all three species since divergence an estimated 6.2 to 8.4 million years BP (Chen and Li 2001), a time scale considerably longer than the time since the *Bos taurus – Bos indicus* divergence.

The neutral theory of evolution predicts an average time to fixation for newly arising neutral alleles of $4N_e$ generations, where N_e is effective population size. Effective population size at the time of the human-chimpanzee split has been estimated as approximately 5,0000 to 100,000 (Chen and Li 2001), although human N_e is estimated to have been only 10,000 over the last 1 to 2 million years (Sherry *et al.* 1997). This means that the average fixation time for a newly arising allele in human would have been of the order of 40,000 generations, or approximately 600,000 to 1 million years, with similar estimates in the case of chimpanzees and gorillas (Ruvolo 1997). That microsatellite polymorphism has been maintained in all three species for almost ten times as long suggests that balancing selection is acting to counter genetic drift.

In the case of cattle, the effective population size is uncertain and varies greatly between breeds. However, estimates for a number of north European breeds indicate values between approximately 30 and 300 (Kantanen *et al.* 1999). Assuming a maximum value for N_e of 1000 and a generation time of approximately 5 years, we would expect an average time to fixation for neutral alleles of as little as 20,000 years. The results obtained here, however, suggest that MHC microsatellite polymorphism may have been retained since the zebu-taurine divergence, which is estimated to have occurred between 117,000 and 850,000 years ago (Bradley *et al.* 1996; MacHugh *et al.* 1997). This again would suggest that balancing selection is acting to limit allele loss through drift.

The results obtained for genetic differentiation of distant populations appear to agree with theoretical predictions and simulation results concerning the effects of balancing selection. Charlesworth *et al.* have derived the result that genetic diversity within populations will increase under balancing selection, whilst the genetic diversity between populations remains approximately constant (Charlesworth *et al.* 1997). Simulations of subdivided populations subject to balancing selection confirm that most of the genetic variation is partitioned within populations, and relatively little between populations (Charlesworth *et al.* 1997). An increase in intra-population diversity would decrease the probability of gene identity for two genes drawn from the same population, and thereby lead to a decrease in Nei's D_S distance (**Equation 2.8**), as observed in this study for zebu-taurine distances calculated using MHC-linked loci. What is not clear is whether the observed decrease in D_S is indeed due solely to increased intra-population diversity. Results for the SMM distance $(\delta\mu)^2$ also show reduced zebu-taurine distance for MHC-linked loci. $(\delta\mu)^2$ is dependent on the difference in mean numbers of repeats between populations and does not take account of intra-population diversity (**Equation 2.9**). This suggests that inter-population divergence is actually restrained at MHC-linked loci.

2.4.5 Evidence of selection at gene-linked microsatellites

Theory suggests that microsatellites linked to genes may be affected by selection. The mode of selection is unlikely to be balancing selection, which appears a rare phenomenon. Rather, we expect genes to be under some form of directional selection that will give effects that are essentially opposite to those seen for microsatellites linked to the MHC.

2.4.6 Reduced allelic diversity for gene-linked microsatellites

In contrast to the MHC-linked loci, the gene-linked loci show low allelic diversity. Both the number of distinct alleles at each locus (**Table 2.8**) and heterozygosity (**Table 2.9**) are significantly lower for gene-lined loci than for either MHC-linked or anonymous loci. The proliferation of novel alleles at microsatellites appears to be restrained by selection at closely linked genes. Whereas diversity at MHC-linked microsatellites appears to be enhanced through local overdominant selection, it appears that the reduced diversity at gene-linked microsatellites is due to a different local mode of selection such as homozygote advantage or purifying selection to remove deleterious alleles.

Comparable studies of diversity at neutral loci linked to genes under selection using wild populations are relatively rare. Indirect evidence suggesting that selection at genes influences microsatellite diversity is provided by studies correlating environmental factors and microsatellite diversity in Israeli wheat and barley populations. Mean heterozygosity across multiple microsatellites is higher for wild emmer wheat populations from drier areas than for populations from wetter areas (Li *et al.* 2000b). Somewhat counter-intuitively, (and in contrast with the results for bovine gene-linked loci reported here), it is suggested that selection for aridity-tolerance causes elevated gene diversity. This may be caused by impaired DNA repair or increased replication error due to physiological stress. Such a mechanism would not explain differences between gene-linked and anonymous loci as observed here for cattle. More relevantly, gene diversity at individual microsatellite loci in wild emmer wheat correlates with soil type. For some loci, populations growing in basalt soils show significantly higher diversity than populations growing in terra-rossa soils, with the situation reversed at other loci (Li *et al.* 2000a). Similar results are found for Israeli wild barley populations, for which gene diversity at certain individual microsatellite loci is correlated with a variety of environmental factors (Turpeinen *et al.* 2001). These results

might be due to selection of different sets of genes in each different environmental circumstance, with consequent effects on diversity for linked microsatellites.

Some more direct evidence for the effects of selection on diversity at linked loci is available. One study of the gene responsible for pyrethroid resistance in tobacco budworm found lower nucleotide diversity in budworm populations exposed to high levels of pesticide and thereby selected for resistance (Taylor *et al.* 1995). In this case, it is unclear whether the observed allelic variation occurs at nucleotide positions directly responsible for the resistance phenotype or at neutral residues. However, similar selection acting in rat populations selected for resistance to the rodenticide warfarin has led to reduced diversity at microsatellites linked to the resistance locus (Kohn *et al.* 2000). Mean gene diversity for 26 microsatellites in a 32 cM region including the resistance locus was calculated for a number of populations subjected to differing levels of warfarin selection, and found to be negatively correlated with the intensity of selection. The proportion of microsatellites showing significant departure from Hardy-Weinberg equilibrium (HWE) was positively correlated with selection intensity, with almost all departures from HWE due to a deficiency of heterozygotes. This suggests homozygote advantage or deleterious allele selection. In this study, we do not observe similar departures from HWE for gene-linked loci, but this may be due to the relative weakness of selection at the genes studied here. In contrast, warfarin causes approximately 95% fatality in rats, implying that selection is very intense.

2.4.7 Theoretical predictions concerning diversity at loci linked to genes

There are two principal models of selection at a gene that could account for the reduced diversity at linked loci: 'hitchhiking', where fixation of a selectively advantageous allele occurs at the gene (Smith and Haigh 1974), and 'background selection', where selection

acts to remove deleterious alleles (Charlesworth *et al.* 1993). In the case of hitchhiking, the allele at the linked neutral locus that is on the same chromosome as the selected allele will increase in frequency, while alleles on chromosomes that are not selectively favoured will decrease in frequency (Smith and Haigh 1974). In the extreme case of complete linkage between selected and neutral loci, only one allele will remain at the neutral locus. For microsatellites, though, the high mutation rate acts to maintain variation. Theory predicts that the ratio of microsatellite repeat size before a hitchhiking event to that after is given by the formula:

$V(t_1)/V(t_0) = 1 - \varepsilon^{8\mu/s}$ (Wiehe 1998)

Equation 2.11

where $V(t_0)$ is initial variance in repeat size, $V(t_1)$ is variance in repeat size at the end of the hitchhiking event, ε is the initial frequency of the selected allele, μ is the mutation rate at the linked microsatellite and s is the selection coefficient at the selected locus. Calculations using **Equation 2.11** (and assuming relatively weak selection) indicate that mutation rates must be of the order of 10^{-4} mutations per generation for the reduction in allele size variance to be detectable, even immediately after the event (Wiehe 1998). In reality, some direct estimates suggest mutation rates for microsatellites a hundred-fold higher (Weber and Wong 1993). Additionally mutation will continue to act after the hitchhiking event to restore variability at the microsatellite locus to its equilibrium value (Slatkin 1995a). The duration of a hitchhiking event is also predicted to be relatively short. The number of generations can be calculated using the formula:

$$t = \frac{-2}{s}\log(\varepsilon)$$

Equation 2.12

where t is time in generations for fixation of the selected allele, ε is the initial frequency of the selected allele, s is the selection coefficient at the selected locus. Values of ε and s between 1×10^{-3} and 0.1 in **Equation 2.12** give fixation times of the order of tens to

thousands of generations. Detectable reduction in microsatellite diversity therefore appears probable only if selection is intense and recent, or if repeated hitchhiking events occur (with greater frequency than the mutation rate) at the same site (Wiehe 1998).

The background selection model proposes that deleterious mutations continuously arise in genes, and that they are selected against through selective disadvantage of the homozygous mutant state (Charlesworth *et al.* 1993). Linked neutral alleles will be co-eliminated along with the deleterious mutations meaning that background selection reduces local nucleotide diversity. Simulation results confirm that background selection leads to reduced intra-population genetic diversity (Charlesworth *et al.* 1993; Charlesworth *et al.* 1997). As background selection is a continual process, there appears to be no requirement for a recent selective event for its effects to be detected. Hence this model of selection appears more plausible in the light of our data than hitchhiking, which would seem to invoke recent hitchhiking events at most or all of the eight gene-linked loci studied.

2.4.8 Skewed allele frequency distributions for gene-linked microsatellites

The gene-linked microsatellites show more skewed (uneven) allele frequency distributions than either the MHC-linked or the anonymous loci as measured by the statistic F_{obs}/F_{exp}, where the expected homozygosity F_{obs} is calculated under the IAM (**Table 2.10**). However, the difference between gene-linked and anonymous loci is not significant. It has been noted that F_{obs} is likely to be affected to a much greater degree by homozygote advantage or disadvantage than by 'background' selection acting to purge deleterious alleles (Watterson 1978). If the mode of selection acting at the bovine genes studied here is background selection, then this could account for the relatively modest decrease in F_{obs}/F_{exp} for the gene-linked microsatellites relative to the anonymous microsatellites.

Comparable studies of the effects of selection acting at genes on the shape of the allele frequency distributions of linked neutral loci are very few. However, one such study investigating the effect of selection at two human genes on polymorphism at adjacent RFLP sites did give similar findings. The loci studied were HRAS-1, tightly linked to the Harvey-ras oncogene and the insulin gene, and D14S1, linked to the heavy-chain immunoglobulin. The modes of selection acting are thought to be different, with heterozygote advantage acting at D14S1 and heterozygote disadvantage or deleterious-allele selection acting at HRAS-1. Comparisons of observed RFLP homozygosity values for three human populations with expected homozygosity values obtained under the Ewens' sampling distribution showed an excess of homozygosity at HRAS-1 and a deficiency at D14S1 (Clark 1987).

2.4.9 Low inter-population genetic distance at gene-linked microsatellites

Both infinite allele model and stepwise mutation model genetic distances between relatively distant populations calculated using gene-linked microsatellites are lower than distances calculated using anonymous microsatellites, although not as low as distances calculated for MHC-linked microsatellites (**Table 2.11**).

The evidence from the studies of neutral loci linked to genes under (non-balancing) selection discussed suggests that inter-population genetic differentiation is actually increased by selection. For the tobacco budworm populations selected for insecticide resistance, inter-population differentiation as measured by F_{ST} was greater at the locus encoding resistance than at unlinked loci (Taylor *et al.* 1995). Similarly, in case of the rat populations selected for warfarin resistance, maximal allelic differentiation between warfarin-resistant and susceptible populations occurred in close proximity to the resistance locus (Kohn *et al.* 2000). One significant point is that the populations experienced widely

varying levels of selection due to different treatment regimens. It would be interesting to examine the effects on genetic differentiation of equivalent levels of selection acting in different populations. The relative levels of selection acting on the cattle genes studied here are unknown, although there is no reason to believe that population-specific selection has occurred at any of the genes.

Theory shows that if different alleles at a gene are selected in two sub-populations, inter-population differentiation will increase at a linked locus (Charlesworth *et al.* 1997). This is presumably because the neutral allele hitchhiking to high frequency is different in each sub-population. Inter-population differentiation at neutral loci may even increase when the same allele is selected in different populations. Slatkin and Wiehe (1998) assume that a selectively advantageous allele at a gene arising in one sub-population goes to fixation, and is then carried to another sub-population by migration of individuals. If a migrant carries a selected allele on the same chromosome as a low frequency allele at a linked neutral locus, then this low-frequency allele is likely to be driven to a high frequency in the new sub-population, thus increasing inter-population differentiation. The effect is most pronounced if the number of migrants is very small ($2Nm << 1$ where N is population size and m the migration rate per generation) (Slatkin and Wiehe 1998). These theoretical findings appear to conflict with the idea that the observed reduction in inter-population distance at gene-linked loci is due to hitchhiking. Background selection, though, may be able to account for the reduced inter-population genetic distance. Theory predicts that only intra-population diversity at neutral loci will be reduced by background selection. Simulation results, however, show that inter-population differentiation is also reduced (Charlesworth *et al.* 1997). The reduction is particularly extreme for populations with a high degree of inbreeding. This may apply more in the case of cattle than for example in humans due to the relatively high variance in male reproductive success in cattle, which leads to a low

effective population size. Why background selection should reduce inter-population diversity is unclear. One suggestion advanced by Charlesworth *et al.* (1997) is heterosis occurring due to migration between two populations, each with high levels of linked deleterious mutations. Hybrid individuals resulting from crosses between parents from each population may be fitter than individuals resulting from within-population matings. Enhanced hybrid fitness would thus put a brake on inter-population differentiation. In the case of cattle, zebu breeds are known to have some taurine genetic input, and even the purest African breeds have a very small zebu component (MacHugh *et al.* 1997) so hybrid heterosis might have some validity. However, further simulation by Charlesworth *et al.* (1997) in which the dominance coefficients of background mutations was increased, thus theoretically reducing the heterotic fitness advantage of hybrids, did not dramatically increase inter-population differentiation. The authors comment that further research will be required to understand the phenomenon.

2.4.10 Range of selective effects

The effects of selection on neutral markers will be strongest when markers are very tightly linked to the selected locus, and will diminish with increasing recombination between the loci. Indications from theory and simulations are that the effects of balancing selection or hitchhiking on a neutral locus will be greatest when there is no recombination between selected and neutral loci. Selective effects will remain significant for neutral loci 1 to 2cM distant, but will be practically undetectable for neutral loci at 10cM or more from the selected locus (Charlesworth *et al.* 1997). Analysis of background selection has tended to consider the combined effects of selection at multiple loci distributed along a chromosome, rather than individual loci (Hudson and Kaplan 1995; Nordborg *et al.* 1996). Consequently, estimates of the range of selective influence for individual loci subject to background selection are unclear. Rather, studies have focussed on the effects of varying

recombination rate on regional diversity, demonstrating that the effects of background selection are likely to be greatest in regions of suppressed recombination (Hudson and Kaplan 1995; Nordborg *et al.* 1996).

The MHC loci studied are very closely genetically linked (Kappes *et al.* 1997). All five are within 3cM (**Figure 2.7**), although the physical map indicates that locus TAMLS113.3 is quite distant (17 centi-rays) from the other four loci, which are all within five centi-rays (Band *et al.* 1998). The nearest markers outside the MHC are PRL (5.2cM distally) and RM185 (7.1cM distally) and BM1258 (11.9cM proximally). PRL is very closely linked to the prolactin gene, and in terms of the population genetic parameters analysed shows complete contrast to the MHC-linked loci (**Figure 2.7**). It has the lowest heterozygosity and number of alleles per locus of all 67 loci, and the third lowest value of F_{obs}/F_{exp}, indicating a highly skewed allele distribution. Locus RM185 ranks 11th highest for heterozygosity and 12th highest for the number of distinct alleles. The genetic distance between zebu and taurine breeds measured using RM185 is also relatively low (14th lowest of 67 for D_S (**Figure 2.7**), 20th lowest of 59 for $(\delta\mu)^2$). These results suggest two things: first that RM185 may be influenced by the effects of selection at the MHC, despite being at a distance of over 7cM(**Figure 2.7**); second that RM185 does not appear to be influenced by selection at the nearby prolactin gene which lies between RM185 and the MHC.

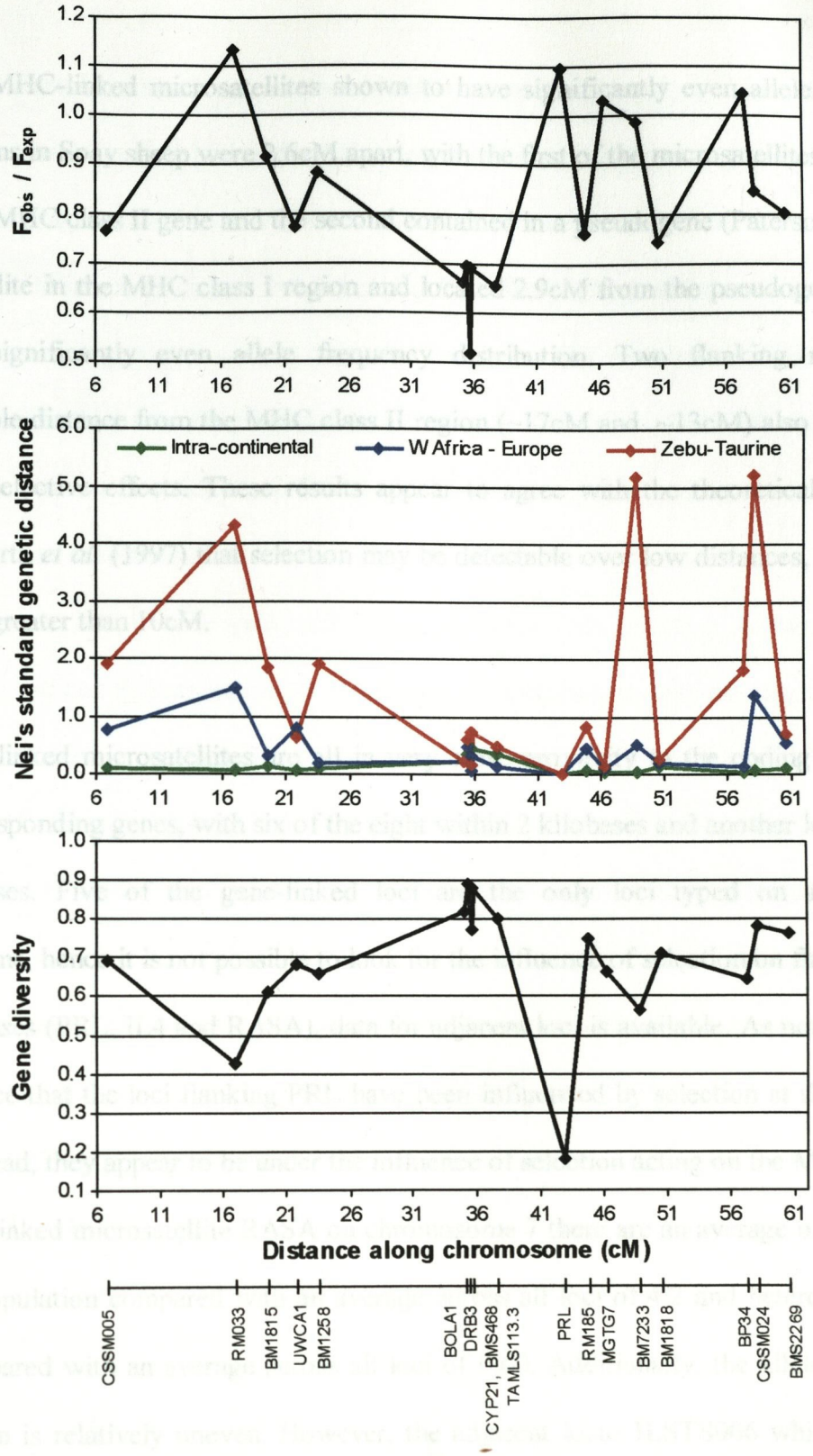

Figure 2.7: Departure from MDE, inter-population genetic distances and gene diversity for 18 microsatellite loci on BTA23

Values of F_{obs}/F_{exp} and gene diversity are averages across populations. For genetic distances, 'intra-continental' distances are averages of the three pairwise intra-continental distances; W Africa – Europe distances are averages of the four pairwise African taurine – European taurine distances; Zebu-taurine distances are the averages of the eight pairwise *B. indicus* - *B. taurus* distances. Locus positions are from Kappes *et al.* (1997), except for DRB3, which is placed between Bola1 and CYP21 as in van Eijk (1995) and MGTG7, for which the position is estimated from Heyen *et al.* (1999)

The two MHC-linked microsatellites shown to have significantly even allele frequency distributions in Soay sheep were 2.6cM apart, with the first of the microsatellites contained within an MHC class II gene and the second contained in a pseudogene (Paterson 1998). A microsatellite in the MHC class I region and located 2.9cM from the pseudogene did not show a significantly even allele frequency distribution. Two flanking markers at considerable distance from the MHC class II region (~17cM and ~13cM) also showed no apparent selective effects. These results appear to agree with the theoretical results of Charlesworth *et al.* (1997) that selection may be detectable over low distances, but not for distances greater than 10cM.

The gene-linked microsatellites are all in very close proximity to the coding regions of their corresponding genes, with six of the eight within 2 kilobases and another locus within 35 kilobases. Five of the gene-linked loci are the only loci typed on a particular chromosome, hence it is not possible to look for the influence of selection on flanking loci. In three cases (PRL, IL4 and RASA), data for adjacent loci is available. As noted, there is no evidence that the loci flanking PRL have been influenced by selection at the prolactin gene. Instead, they appear to be under the influence of selection acting on the MHC region. For gene-linked microsatellite RASA on chromosome 7 there are an average of 2.8 alleles in each population compared with an average across all loci of 4.2 and heterozygosity is 0.50 compared with an average across all loci of 0.60. Additionally, the allele frequency distribution is relatively uneven. However, the adjacent locus ILSTS006 which is 13cM away shows above-average diversity (4.5 alleles, heterozygosity of 0.71) and an even allele distribution. Finally, gene-linked microsatellite IL4 is of average diversity (4.0 alleles, heterozygosity of 0.58), hence it is not realistic to expect the adjacent loci typed to show selective influence due to proximity to IL4.

Loci under directional selection, whether hitchhiking or background selection, are probably much more common than loci under balancing selection. The balancing selection seen at the MHC appears something of a special case. Thus we would expect more microsatellites to show the influence of directional selection. The clear difference between known gene-linked and anonymous loci seen in this study suggests that most microsatellites are not greatly influenced by selection. Taking the estimate of 30,000 genes in 3100 megabases for the human genome (Lander *et al.* 2001), we can suppose that inter-genic separation in mammals is of the order of 100 kilobases. Assuming that microsatellites are indeed on average within 100kb of a gene, this suggests either that selection at most genes is relatively weak and thus undetectable at linked microsatellites, or that the effects are short-range, and so much more likely to be detected over distances of the order of 1 kb, as for the loci studied here, than 100 kb. Studying microsatellites at a range of intervals from genes will be useful in determining the scale over which selection is indeed detectable.

2.4.11 Mutation rate variability as a possible explanation of observations

There is a possibility that the differences in population-genetic behaviour between gene-linked, MHC-linked and anonymous loci might be due to differences in mutation rate between microsatellite loci. Higher mutation rates could cause increased locus diversity and inter-population genetic distance. It is less clear, though, that higher mutation rates would affect the skew of allele frequency distributions, as seen in this study.

There is evidence that mutation rate varies between microsatellite loci, possibly by several orders of magnitude (Brinkmann *et al.* 1998; Harr *et al.* 1998). Mutation rates may even vary considerably between alleles at the same locus (Jin *et al.* 1996; Schlotterer *et al.* 1998). A number of sources of variation in mutation rate have been identified. Relevant

factors in this study are the identity of the repeat units in the microsatellite array, the number of repeat units in the array, and the presence of interruptions in the array. Comparison of variability of microsatellites consisting of GT/CA repeats, TC/GA repeats and AT/TA repeats in *Drosophila melanogaster* suggests that GT/CA repeats evolve fastest and AT/TA repeats slowest (Bachtrog *et al.* 2000). Some evidence points to higher mutation rates for microsatellites with larger numbers of repeat units (Brinkmann *et al.* 1998; Goldstein and Clark 1995), although other studies have failed to demonstrate any association (Valdes *et al.* 1993). Mutation rates may also be higher for perfect arrays of tandem repeats (e.g.$[CA]_{20}$) than for arrays which are interrupted (e.g.$[CA]_{10}[CT][CA]_9$) (Brinkmann *et al.* 1998; Goldstein and Clark 1995).

Sequences of the alleles observed in this study were not obtained. Published repeat sequences are available for 58 of the 67 loci, including all gene-linked and MHC-linked loci. Unfortunately, though, the relationship of the published alleles to the alleles observed in this study cannot be confirmed without access to the same reference samples. Nevertheless, comparison of the published repeat motifs (where available) for the gene-linked and anonymous loci (**Table 2.2**) does not reveal any striking differences in repeat composition or length. Most loci contain primarily CA/GT repeat units. This applies to all five MHC-linked loci, seven of the eight gene-linked microsatellites and 42 of the 45 anonymous loci for which sequences are available. A large number of loci show complex structures including compound repeats of different dinucleotides or interruptions to repeat arrays. Three of the five MHC loci have complex structures, as do four of the eight gene-linked loci and 22 of the 44 sequenced anonymous loci. Finally, lengths of non-interrupted repeat arrays do not differ obviously between the classes of loci. While chance differences in mutation rate between the loci in the three different classes cannot be discounted as an

explanation for some of the phenomena described, we believe that selection is the more likely cause.

2.4.12 Effects of selection on loci evolving under the SMM

We have compared observed homozygosity values for the population-locus combinations typed with corresponding MDE expected values estimated under both the infinite allele and stepwise mutation models. Infinite allele estimates were obtained using Ewens' sampling distribution for MDE alleles (Ewens 1972). SMM estimates were obtained using the simulation method of Cornuet and Luikart (1996).

In the case of the SMM, we find a highly significant correlation between the number of distinct alleles and estimated expected homozygosity. Interestingly, Cornuet and Luikart (1996) find that the magnitude of heterozygosity excess following a population bottleneck declines much more rapidly for SMM loci than for IAM loci as the number of alleles increases. In some cases, heterozygosity deficiency is seen at SMM loci with higher numbers of alleles, even though population bottlenecks are predicted to cause a transient increase in heterozygosity (Nei *et al.* 1975). However, it must be noted that Cornuet and Luikart do not test IAM loci using SMM estimates of heterozygosity or vice versa, hence it is unclear from their results whether the IAM estimates would prove a less biased indicator of departure from MDE for microsatellites with high variance in allele number, as we suspect. Cornuet and Luikart do, though, find that the probability of detecting heterozygosity excess for SMM loci decreases as locus heterozygosity increases. In the light of these and our results, the SMM estimates of expected homozygosity were not used in testing for departure from MDE.

The IAM estimates of expected homozygosity at MDE are consistently lower than observed homozygosity, both for the real dataset (**Table 2.10**), and for simulated hybrid populations obtained by sampling alleles from this dataset (**Figure 2.6**). Of 393 polymorphic population-locus combinations, expected heterozygosity exceeds the IAM estimate in 315 cases. In contrast, expected heterozygosity exceeds the SMM estimate in just 155 cases. This result is not unexpected, given that the actual mode of mutation for the loci studied is likely to be intermediate between the IAM and SMM. Mutation at microsatellites is more likely to result in a pre-existing allele, thus limiting the number of rare alleles and causing apparent heterozygosity excess given the number of alleles. Cornuet and Luikart (1996) test for heterozygosity excess or deficiency at microsatellite loci using both IAM and SMM neutral estimates in four populations of different demographic history. For the only population for which the IAM and SMM estimates give inconsistent results, observed heterozygosity exceeds the SMM estimate at nine loci and the IAM estimate at all 14 loci. This supports our finding that IAM estimates of heterozygosity under neutrality tend to exceed observed heterozygosity for microsatellites.

These results suggest that caution should be exercised in inferring departure from MDE at individual microsatellite loci. Tests such as those of Watterson (Watterson 1978) or Slatkin (Slatkin 1994) which use Ewens' sampling distribution, with its assumption of infinite allele mutation, seem biased towards indicating heterozygosity excess. Equally, tests using estimates of expected heterozygosity under a strict SMM model may be unreliable if there is a large variance in the number of alleles at different loci. We have sought to overcome these problems by assuming the IAM to estimate expected heterozygosity, and then by performing inter-locus comparison of relative heterozygosity excess or deficiency. Alternatively, Cornuet and Luikart (1996) have proposed using a mutation model intermediate between the IAM an SMM to estimate expected heterozygosity, and this may

prove useful if the relative proportion of stepwise to non-stepwise mutations can be estimated.

2.4.13 Effects of admixture on allele distributions

It has previously been reported that heterozygosity deficiency can result from admixture due to the introduction of low frequency rare alleles (Chakraborty 1990). Chakraborty showed that pooling samples from a number of Asian human populations gave a hybrid population in which the number of alleles exceeded that expected for a population at MDE, given the level of heterozygosity. Here we find that admixture can also have the opposite effect. When samples from different populations are pooled, the resulting simulated hybrid population can tend either towards heterozygosity deficiency or excess, relative to the parental populations. The direction of the departure from MDE is determined by the evolutionary separation of the parental populations and the admixture proportion (**Figure 2.6**). For hybrid populations constructed from the highly divergent zebu and taurine populations, there is an increase in observed relative to expected homozygosity (i.e. tendency to heterozygosity deficiency) if the relative genetic contributions from the two parental populations are uneven ($\geq$80% genes from parental population 1, $\leq$ 20% genes from parental population 2). As the contributions from the parental populations become more even, there is a decrease in observed relative to expected homozygosity (tendency to heterozygosity excess) which is greatest when parental population contributions are equal. The latter finding presumably reflects the fact that zebu and taurine allele frequency distributions tend to be very different, so an approximately equal mixture of the distributions results in a bi- or multi-modal distribution, which will tend towards heterozygosity excess.

In contrast, when the more closely related European and West African taurine breeds are used to construct hybrid populations, there is no increase in observed relative to expected homozygosity when parental population contributions are approximately equal (i.e. no tendency to heterozygosity excess). This reflects the similarity in allele frequency distributions in the parental populations. Closely related populations are unlikely to differ in the modal allele, hence admixed populations are unlikely to be bi- or multi-modal. There is, however, an increase in heterozygosity deficiency when a small proportion of European genes are introduced into a predominantly West African background. This effect is not seen when West African genes are introduced at low frequency into a European background. This is likely to be due to the low level of genetic diversity in West African breeds. The West African breeds have a mean heterozygosity across loci of 0.57 and a standardised mean number of alleles per locus of 5.14 (average of 100 resamplings of 56 alleles from each breed at each locus) whereas the European breeds have a mean heterozygosity of 0.63 and a standardised mean number of alleles per locus of 5.28. It is therefore more likely that European alleles introduced into an African background will be novel alleles not previously present than for African alleles introduced into a European background. The results for admixture between European and African taurine cattle agree with the results of Chakraborty (1990) for admixed populations constructed from human populations that are considerably evolutionarily closer than zebu and taurine cattle.

These results underline the importance of considering all possible causes of departure from MDE at loci. It would be unwise to infer selection from heterozygosity excess or deficiency without considering the possibility of admixture, or indeed of other causes such as population bottlenecks (Cornuet and Luikart 1996). These factors should affect all loci, hence inter-locus comparison using many loci, as in this study, should reveal which individual loci have been subject to selection.

2.4.14 Conclusions

Identification of the footprints of selection in specific genomic regions through population genetic analysis of diversity of mapped microsatellites markers appears feasible. Such analysis may be a useful complement to pedigree-based gene mapping, and may also be applied where no pedigree information is available. It is apparent that a number of microsatellite loci must be typed, as there is considerable stochastic variation in allele spectra between loci. Furthermore, it appears that microsatellite loci will only be significantly influenced by selection at neighbouring genes if the distance between the selected locus and the microsatellite is small.

CHAPTER 3

APPLICATION OF POPULATION GENETICS TO CANDIDATE CHROMOSOMAL REGIONS FOR TRYPANOTOLERANCE

3.1 INTRODUCTION

In the light of the encouraging results from **Chapter 2**, the same population genetic analyses used in that chapter are applied to three quantitative trait loci (QTL) identified as possibly conferring tolerance to trypanosomiasis in African cattle. Microsatellites across the three regions are investigated for evidence of selection and compared with 'background' loci distributed elsewhere in the genome.

3.1.1 QTL regions identified in trypanotolerance genome scan

As described in **Section 1.2.3**, an F2 pedigree of seven major and several minor families of cross-bred N'Dama-Boran animals was produced at ILRI and monitored for body weight, parasitaemia (parasite load) and anaemia after challenge with trypanosomes. Animals were genotyped for 170-200 microsatellite markers. Data from the 177 animals belonging to the seven major families was analysed to reveal associations between phenotypic traits and marker loci. Several different analyses have been performed using different combinations of traits, varying numbers of animals (due in part to incomplete genotyping of the pedigree) and analytical approaches. Many potential trypanotolerance QTL have been identified, although it is likely that a large number will ultimately turn out to be false positives. Three of the most promising QTL are on chromosomes BTA2, BTA5 and BTA7.

3.1.1a BTA2 trypanotolerance QTL

Markers in the central region of BTA2 are associated with variation in packed red blood cell volume (PCV) post trypanosome challenge. Analysis at the Institute of Evolution, University of Haifa (Soller, unpublished data) considered a number of different traits calculated from the PCV measures. Soller *et al.* found possible QTL at locus ILSTS098 (**Figure 3.3**) governing the overall drop in PCV (unspecified model, Lod 2.79), again at ILSTS098 for the maximum fall in PCV (recessive, Lod 2.16), at locus BM4440 affecting

variance in PCV over challenge (additive, Lod 2.31) and again at ILSTS098 affecting the final PCV value at the end of the challenge period (additive, Lod 2.26). Combined trait analysis was also performed using data on the percentage decline in body weight over the entire challenge period and each of the above PCV traits in turn. This again identified locus BM4440 as a potential QTL, with a much higher maximal Lod score of 5.1. Independent Analysis at ILRI indicated an additive N'Dama-derived QTL affecting the percentage decline in PCV from initial value to minimum at locus BMS803 with a Lod score of 3.23. The position of this QTL is shown in **Figure 3.3**.

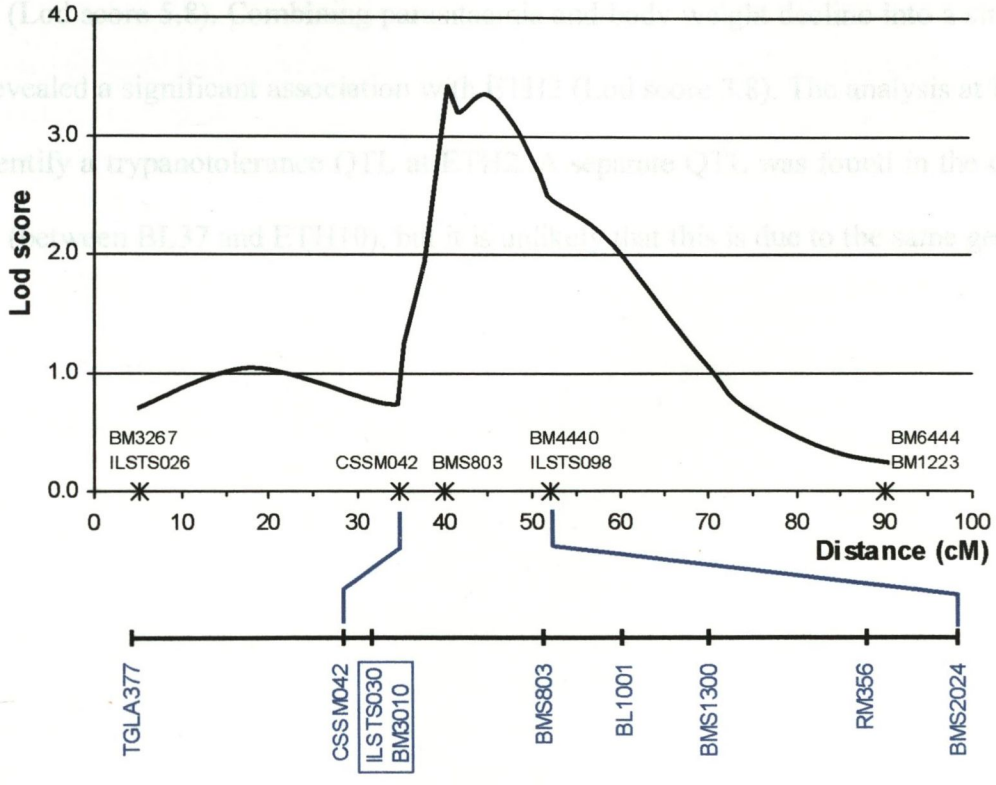

Figure 3.1: Position of QTL for PCV decrease on BTA2 and of loci studied in field populations

Lod scores taken from Hanotte *et al.*, in preparation. Microsatellite loci in black type were used in the ILRI trypanotolerance gene scan. Positions of these loci are taken from Hanotte *et al.*, in preparation. Loci in blue type were typed in field populations in this study. Positions are as in Kappes *et al.* (1997).

3.1.1b BTA5 trypanotolerance QTL

Markers towards the telomeric end of BTA5 are associated with PCV traits. Analysis at the University of Haifa (Soller, unpublished data) shows an association between locus ETH2 (**Figure 3.2**) and the overall decrease in PCV. If the QTL is recessive, the Lod score is 3.07, and if the dominance model is unrestricted the Lod score is 3.1. Lower Lod scores (below the threshold of 3.0) are found for associations between ETH2 and the related PCV traits outlined in **Section 3.1.1a**, with the best score in each case obtained under the assumption that the QTL is recessive. Combined trait analysis using parasitaemia data and PCV data indicates a highly significant association between the combined trait and locus ETH2 (Lod score 5.8). Combining parasitaemia and body weight decline into a single trait also revealed a significant association with ETH2 (Lod score 3.8). The analysis at ILRI did not identify a trypanotolerance QTL at ETH2. A separate QTL was found in the centre of BTA5 (between BL37 and ETH10), but it is unlikely that this is due to the same gene.

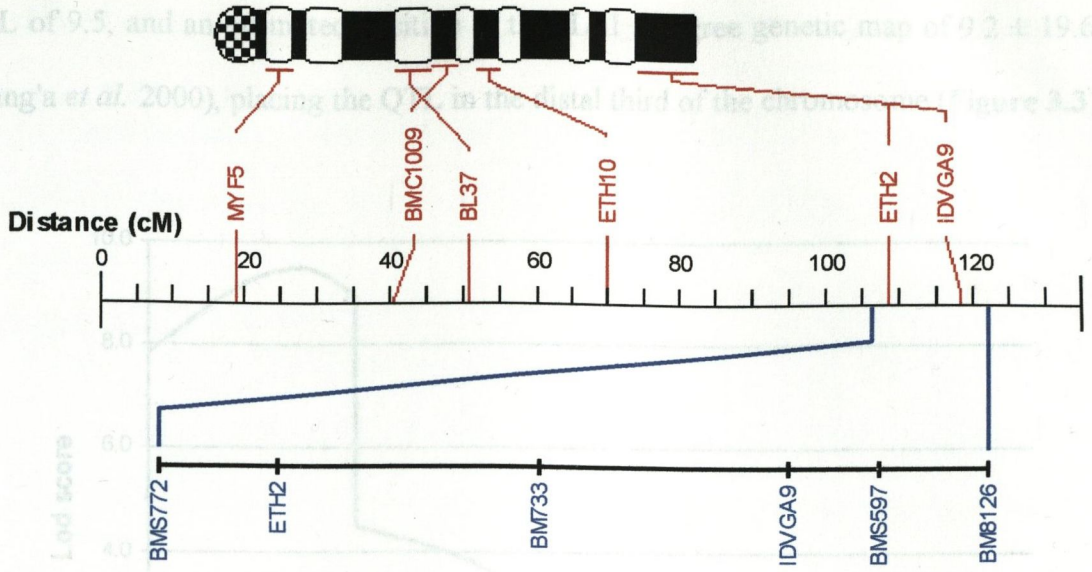

Figure 3.2: Position of QTL for PCV decrease on BTA5 and of loci studied in field populations

Upper part of figure shows cytogenetic banding pattern of BTA5 and loci (in red type) mapped cytogenetically and by genetic linkage (adapted from Kappes *et al.* (1997)). Loci in blue type were typed in field populations in this study. Locus positions are as in Kappes *et al.* (1997).

3.1.1c BTA7 trypanotolerance QTL

Markers towards the centromeric end of BTA7 are significantly associated with two different trypanotolerance traits. Analysis at ILRI (Hanotte *et al.*, in preparation) and the University of Haifa (Soller, unpublished data) revealed Lod scores of 2.9 and 3.5 respectively for a recessive N'Dama-derived QTL for low parasitaemia at locus BP41 (**Figure 3.3**). This QTL accounts for 11% of the observed phenotypic variation. In the same region, from locus BP41 to RM006, there is Lod score of 2.9 for an N'Dama-derived dominant QTL for percentage decrease in PCV during trypanosome infection accounting for 11% of observed variation (Hanotte *et al.*, in preparation). Analysis carried out at Wageningen University (van der Waaij and van Arendonk, unpublished data) found a weak association between loci on BTA7 and a similar PCV trait, supporting this second result. Dual-trait analysis carried out under the assumption that the same gene is responsible for both parasitaemia and PCV effects gives a Lod score for the combined

QTL of 9.5, and an estimated position in the ILRI pedigree genetic map of 9.2 ± 19.6cM (Kang'a *et al.* 2000), placing the QTL in the distal third of the chromosome (**Figure 3.3**).

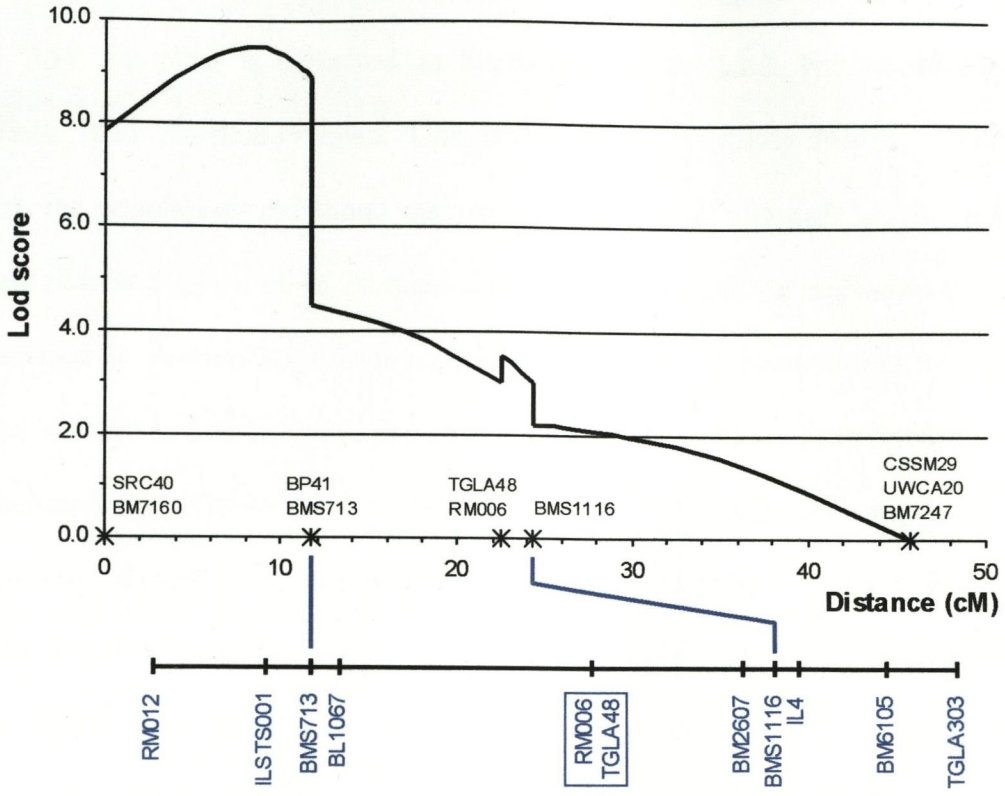

Figure 3.3: Position of QTL for parasitaemia and PCV decrease on BTA7 loci and of loci studied in field populations

Lod scores taken from Kang'a *et al.* (2000). Microsatellite loci in black type were used in the ILRI trypanotolerance gene scan. Positions of these loci are taken from Hanotte *et al.*, in preparation. Loci in blue type were typed in field populations in this study. Positions are as in Gu *et al.* (2000).

3.1.2 Trypanotolerance QTL microsatellite expectations

Results from **Chapter 2** suggest that microsatellites in the vicinity of the trypanotolerance QTL may be influenced by selection. As discussed above, analysis suggests that the BTA2 QTL acts recessively or additively, that on BTA5 is recessive, and that on BTA7 is dominant or recessive. There is no suggestion of overdominance at any of the QTL, hence we may expect microsatellites linked to the QTL to be affected in comparable ways to the gene-linked loci of **Chapter 2**. We might therefore expect to see reduced allelic diversity

and skewed allele distributions. Effects on inter-population differentiation are, as discussed in **Section 2.4.9** harder to predict.

3.2 MATERIALS AND METHODS

3.2.1 Breeds used in comparison of QTL regions

Samples of eight cattle breeds were used in studying the trypanotolerance QTL regions (**Table 3.1**). Sampling is described in **Sections 2.2.1** to **2.2.3**. For six of the breeds (Charolais, Irish Holstein-Friesian, Guinean N'Dama, Togolese Somba, Nellore and Ongole), the populations are those analysed in **Chapter 2**, although in the case of the Togolese Somba a subset of 45 samples was used. Origins and descriptions of the breeds are described in **Section 2.2.4**.Two additional West African zebu-taurine hybrid breeds (Djakoré and hybrid Somba) were also studied. The Djakoré derive from Senegal and are a cross between zebu Gobra cattle from the relatively arid North of Senegal and taurine N'Dama from further south (**Figure 1.2**). Djakoré are considerably larger than N'Dama (~135cM at withers), and have a hump (though not pronounced) due to their partial zebu ancestry (Trail *et al.* 1979). The hybrid Somba (also known as Borgou) derive from Northern Togo and Benin (**Figure 1.2**). They are a cross between zebu Fulani cattle from further north in Upper Volta and the local taurine Somba. Hybrid Somba cattle show considerable variation in size depending on the percentage of zebu genes (Trail *et al.* 1979).

Breed	Sub-species	Breed origin	Sampling location	Sample size
Charolais	*B. taurus*	France	Ireland	34
Holstein-Friesian	*B. taurus*	Netherlands	Ireland	40
N'Dama	*B. taurus*	W Africa	Guinea	63
Somba	*B. taurus*	W Africa	Northern Togo	45
Ongole	*B. indicus*	S India	S India	30
Nellore	*B. indicus*	S India	Brazil	35
Djakoré	*B. taurus* / *B. indicus* hybrid	W Africa	Senegal	45
Hybrid Somba	*B. taurus* / *B. indicus* hybrid	W Africa	Northern Togo	45

Table 3.1: Details of cattle breeds studied at QTL regions

3.2.2 Microsatellite loci used in comparison of QTL regions

A total of 33 microsatellite loci were typed in the cattle populations studied. Of these, nine are from the QTL region on BTA2, six from the QTL region on BTA5 and eleven from the QTL region on BTA7. The remaining seven loci are 'background' loci distributed throughout the genome (**Table 3.2**). Using the linkage map distances of Kappes *et al.* (1997) for purposes of comparison, markers on BTA2 span 27.8cM, those on BTA5 15.9cM and those on BTA7 30.6cM.

QTL regions				
BTA2	**BTA5**	**BTA7**	**Background loci**	
TGLA377	BMS772	RM012	ETH10	(BTA5)
CSSM042	ETH2	ILSTS001	HEL9	(BTA8)
BM3010*	BM733	BMS713	ETH225	(BTA9)
ILSTS030*	IDVGA9	BL1067	ILSTS005	(BTA10)
BMS803	BMS597	RM006†	INRA005	(BTA12)
BL1001	BM8126	TGLA48†	HEL1	(BTA15)
BMS1300		BM2607	INRA063	(BTA18)
RM356		BMS1116		
BMS2024		IL4		
		BM6105		
		TGLA303		

Table 3.2: Loci studied at three trypanotolerance QTL and 'background' reference loci

QTL loci are listed in chromosomal order. Order of BTA2 loci is from the genetic linkage map of Kappes *et al.* (1997) *BM3010 and ILSTS030 map to the same location. Order for BTA5 loci is also from the genetic linkage map of Kappes *et al.* (1997). Order of BTA7 loci is from the comprehensive linkage map of Gu *et al.* (2000) † RM006 and TGLA48 map to the same location. Chromosomal locations of background loci are given in brackets after the locus names.

Details of primers, repeat motifs and references for the microsatellite loci are given in **Table 2.2**, with the exception of loci IDVGA9 on chromosome BTA5 and BM2607 on chromosome BTA7, for which details are given in **Table 3.3**.

Locus	Chr	Position (cM) *	GenBank No	Forward primer Reverse primer	Repeat motif †	References
IDVGA9	5	118.3	Z27075	GTC AGG TCT AAA CCC AGA GCC TCA AAA GGG CAG AGT TCC AC	$(TG)_{12}$	1, 2
BM2607	7	29.1	G18470	GGC CTG TGA CTC CTT GTA GG TTC CTG TGG GCT GGC TAG	$(GT)_{10}GA(GTG)_{2}$- $N_6(TGTA)_2(TG)_4$	3, 4

Table 3.3: Details of microsatellite loci IDVGA9 and BM2607

'Chr' is the chromosome on which the locus is located. *Position within the chromosome is taken from the on-line MARC bovine genome map http://sol.marc.usda.gov/genome/cattle/cattle.html, originally published as Kappes *et al.* (1997). † Repeat motif is that of the published (GenBank) allele. References: 1 - Ferretti, Leone *et al.* 1994; 2 - Mezzelani, Zhang *et al.* 1995; 3 - Bishop, Kappes *et al.* 1994; 4 - Kappes, Keele *et al.* 1997

3.2.3 QTL Microsatellite PCR and visualisation

PCR optimisation and typing was performed as described in **Sections 2.2.6a** and **2.2.6b**.

PCR products were electrophoresed and visualised as described in **Section 2.2.7**

3.2.4 Analysis

A number of different population-genetic analyses were performed using a variety of different software. Formatting of data for the various software listed was performed using 'the Excel Microsatellite Toolkit' described in **Section 6.1**.

3.2.5 Hardy-Weinberg equilibrium

All polymorphic population-locus combinations were tested for deviation from HWE as described in **Section 2.2.9**.

3.2.6 Intra-population diversity

Expected heterozygosity and mean number of alleles per locus were calculated as described in **Section 2.2.8**.

3.2.7 Inter-population diversity

Nei's standard genetic distance was calculated for all pairwise combinations of populations as described in **Section 2.2.14**.

3.2.8 Admixture estimation for hybrid populations

For a hybrid population formed by admixture of two ancestral populations (designated *H*), the relative genetic input from each of the two ancestral populations (designated *A* and *B*) can be estimated by a least-square method using the formula:

$$\hat{\mu} = \frac{\sum_{l=1}^{L}\sum_{i=1}^{r(l)}\left(P_{Ai}^{(l)} - P_{Bi}^{(l)}\right)\left(P_{Hi}^{(l)} - P_{Ai}^{(l)}\right)/P_{Hi}^{(l)}}{\sum_{l=1}^{L}\sum_{i=1}^{r(l)}\left(P_{Ai}^{(l)} - P_{Bi}^{(l)}\right)^2/P_{Hi}^{(l)}}$$

Equation 3.1

where $\hat{\mu}$ is the proportion of genes from population *A* in the hybrid population and $P_{Ai}^{(l)}$, $P_{Bi}^{(l)}$, and $P_{Hi}^{(l)}$ are the frequencies of the *i*'th allele of $r(l)$ segregating alleles at the *l*'th locus (of *L* loci) in the first and second ancestral populations and the hybrid population respectively (Chakraborty *et al.* 1992).

Variance of $\hat{\mu}$ is given by:

$$V(\hat{\mu}) = MSE \Big/ \sum_{l=1}^{L}\sum_{i=1}^{r(l)}\left(P_{Ai}^{(l)} - P_{Bi}^{(l)}\right)^2/P_{Hi}^{(l)}$$

Equation 3.2

where *MSE* (mean square error) is given by:

$$MSE = \frac{\sum_{l=1}^{L}\sum_{i=1}^{r(l)}\left[\left(P_{Hi}^{(l)} - P_{Bi}^{(l)}\right)\left(\hat{\mu}P_{Ai}^{(l)} - P_{Bi}^{(l)}\right)\right]^2/P_{Hi}^{(l)}}{\left(\sum_{l=1}^{L} r(l)\right) - L}$$

Equation 3.3

Estimation of $\hat{\mu}$ values for the West African hybrid populations studied was performed using the program Admix1.0 (Bertorelle and Excoffier 1998).

3.2.9 Linkage disequilibrium

The non-random association of alleles at different genetic loci into gametes is termed gametic disequilibrium. This can be caused by a variety of factors, one of which is physical linkage, where loci are in proximity on the same chromosome. In this case the term linkage disequilibrium is used. The extent of linkage disequilibrium may be affected by selection. Strong selection for a given allele at a selected locus may, through hitchhiking, drive up frequencies of linked alleles at neighbouring loci, thus increasing the degree of linkage. For each of the three QTL chromosomes, linkage disequilibrium was assessed for all possible pairwise combinations of loci in each population. The tests were carried out using the GenePop program (Raymond and Rousset 1995). The algorithm used is based on analysis of simple contingency tables defined in terms of two-locus genotype counts. It is not necessary to know the gametic phase of animals heterozygous at both loci under consideration. Each contingency table is analysed using a Markov chain method with 50,000 iterations in a manner identical to Fisher's exact test for HWE described in **Section 2.2.9**.

3.3 RESULTS

3.3.1 Tests of Hardy-Weinberg equilibrium

A total of 264 population-locus combinations were typed of which 256 were polymorphic. 16 population-locus combinations deviated from HWE at the 5% level, whereas 12.8 would be expected by chance. At the 1% level, 3 population-locus combinations deviated from HWE, whereas 2.6 would be expected by chance. When a Bonferroni correction is applied to correct for multiple hits, no population-locus combinations show significant deviation from HWE. No loci deviated from HWE at the 5% level in more than one of the eight populations typed, and no populations deviated from HWE at more than three of the 33 loci typed. The results provide no evidence for significant deviation from HWE at any locus or in any population.

3.3.2 Comparison of gene diversity at QTL loci and background loci

When mean gene diversity values (averaged across loci within each QTL region and across background loci) are compared, the highest values in the African breeds, both hybrid and purebred, are those for background loci. Within each of the African breeds, gene diversity values for the QTL regions are consistently lower (**Figure 3.4**). This might be expected if hitchhiking with an adjacent QTL affected the loci. However, the same pattern is seen in Charolais, which have not been subjected to selection for trypanotolerance. Also, in Nellore and Ongole, mean diversity values for BTA5 and BTA7 are lower than for the background loci. Furthermore, gene diversity standard deviations, are high, reflecting substantial variation between individual loci. T tests for each breed only reveal significant differences between gene diversity values for background loci and QTL loci in Djakoré ($p=0.032$) and Ongole ($p=0.015$) but the significance of these results disappear when a Bonferroni correction for multiple testing is applied. Non-parametric (Mann-Whitney) tests

also show no significant difference between QTL loci and background loci in any breed. Separate t tests comparing gene diversity values for each QTL region in turn with those for background loci also reveal no significant differences in the African breeds.

Overall, mean gene diversity is lowest for the two purebred African breeds, with values of 0.47 ± 0.04 for N'Dama and 0.54 ± 0.03 Somba compared with 0.56 to 0.60 for other pure breeds. Gene diversities for the hybrid African breeds are higher than for the corresponding taurine ancestral populations (N'Dama in the case of Djakoré; Somba in the case of hybrid Somba) due to zebu genetic input.

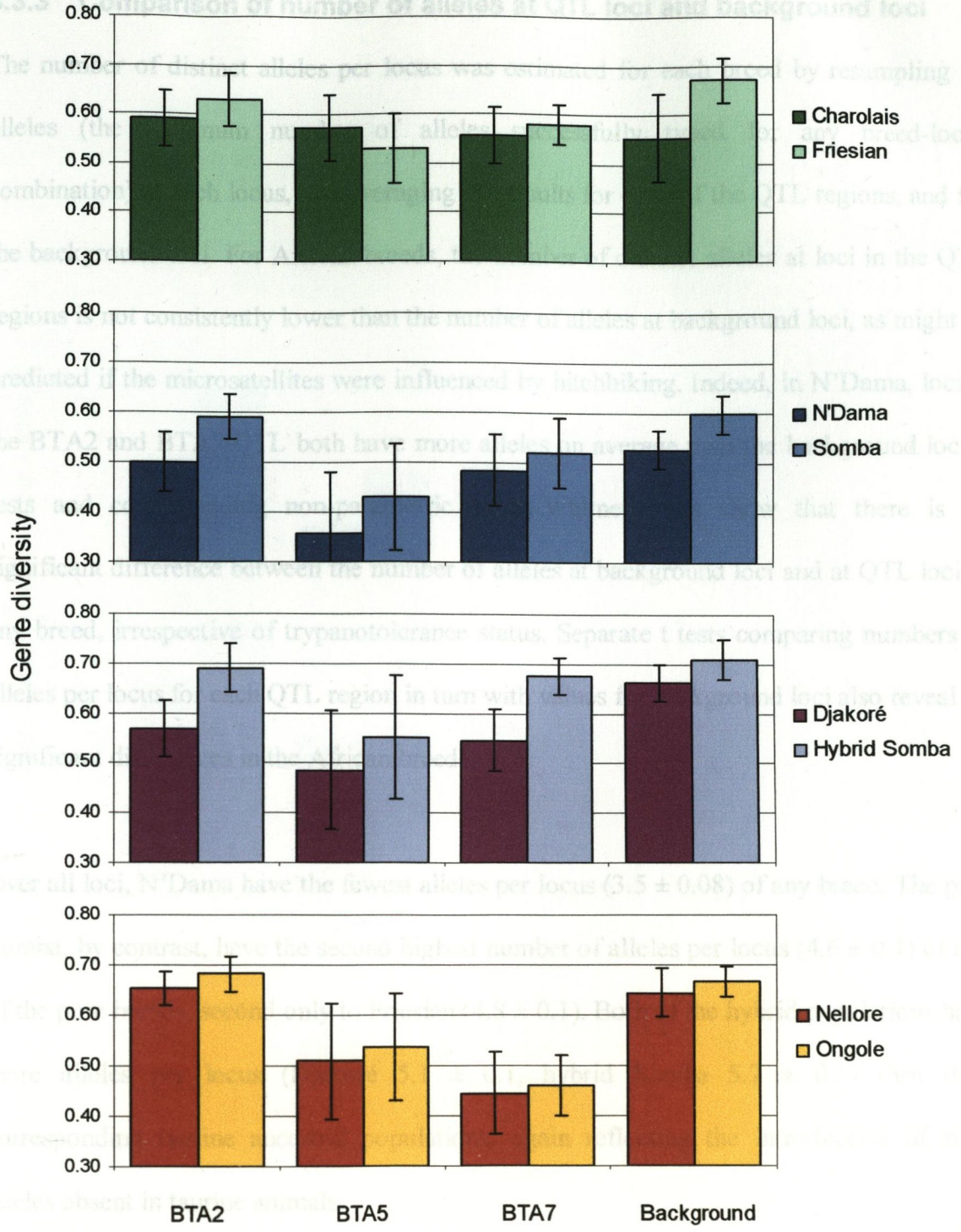

Figure 3.4: Mean gene diversity for three candidate trypanotolerance QTL regions and background loci in trypanotolerant and susceptible cattle breeds

For each QTL region and for the background loci, gene diversity (expected heterozygosity) values shown are averages across loci.

3.3.3 Comparison of number of alleles at QTL loci and background loci

The number of distinct alleles per locus was estimated for each breed by resampling 58 alleles (the minimum number of alleles successfully typed for any breed-locus combination) at each locus, and averaging the results for each of the QTL regions, and for the background loci. For African breeds, the number of distinct alleles at loci in the QTL regions is not consistently lower than the number of alleles at background loci, as might be predicted if the microsatellites were influenced by hitchhiking. Indeed, in N'Dama, loci at the BTA2 and BTA7 QTL both have more alleles on average than the background loci. t tests and corresponding non-parametric Mann-Whitney tests show that there is no significant difference between the number of alleles at background loci and at QTL loci in any breed, irrespective of trypanotolerance status. Separate t tests comparing numbers of alleles per locus for each QTL region in turn with values for background loci also reveal no significant differences in the African breeds.

Over all loci, N'Dama have the fewest alleles per locus (3.5 ± 0.08) of any breed. The pure Somba, by contrast, have the second highest number of alleles per locus (4.6 ± 0.1) of any of the pure breeds, second only to Friesian (4.8 ± 0.1). Both of the hybrid populations have more alleles per locus (Djakoré 5.1 ± 0.1, hybrid Somba 5.7 ± 0.1) than their corresponding taurine ancestral populations, again reflecting the introduction of zebu alleles absent in taurine animals.

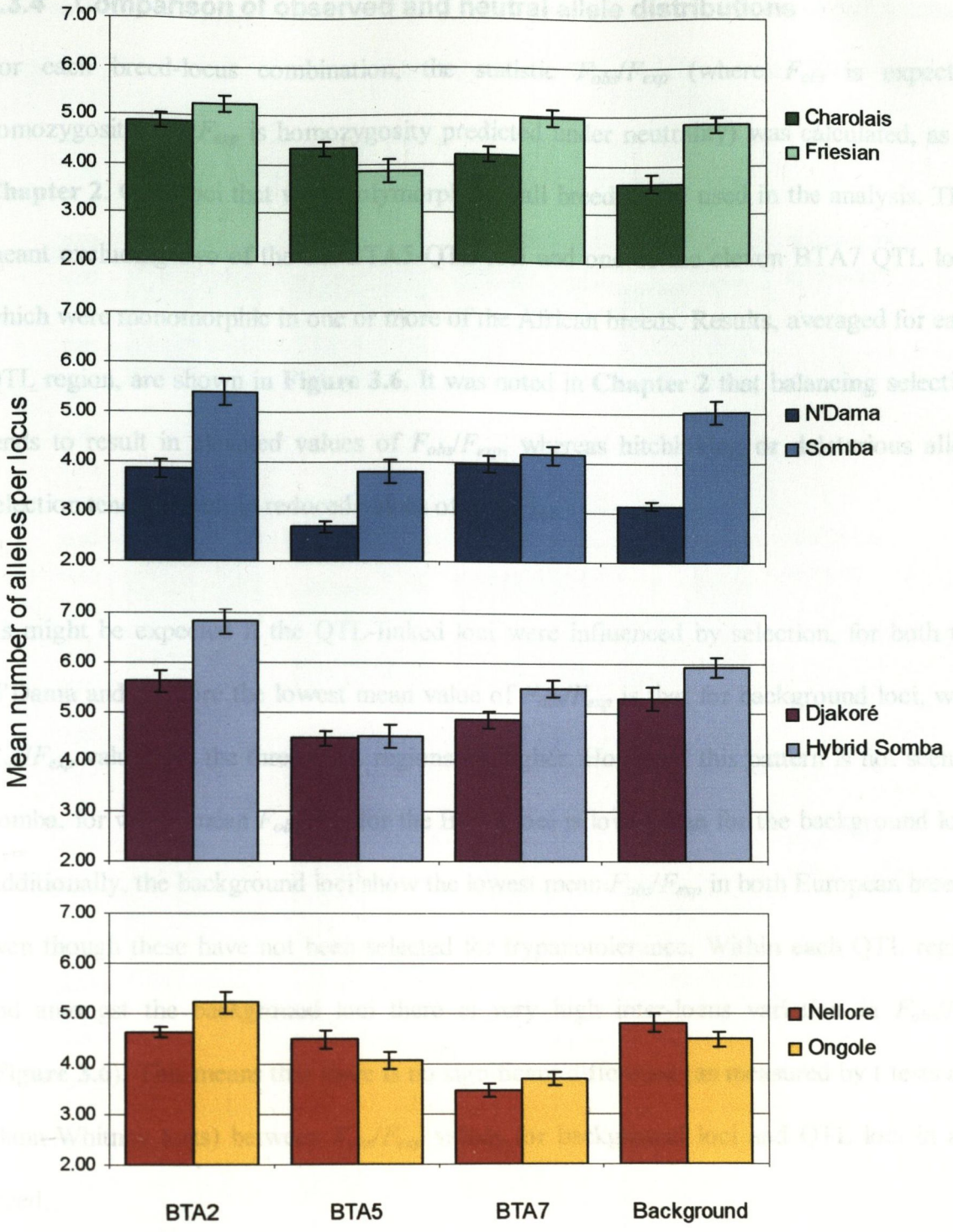

Figure 3.5: Mean number of alleles per locus for three candidate trypanotolerance QTL regions and background loci in trypanotolerant and susceptible cattle breeds

Estimates are averages across loci of 100 replicate resamplings of 58 alleles at each population-locus combination

3.3.4 Comparison of observed and neutral allele distributions

For each breed-locus combination, the statistic F_{obs}/F_{exp} (where F_{obs} is expected homozygosity and F_{exp} is homozygosity predicted under neutrality) was calculated, as in **Chapter 2**. Only loci that were polymorphic in all breeds were used in the analysis. This meant excluding two of the six BTA5 QTL loci and one of the eleven BTA7 QTL loci, which were monomorphic in one or more of the African breeds. Results, averaged for each QTL region, are shown in **Figure 3.6**. It was noted in **Chapter 2** that balancing selection tends to result in elevated values of F_{obs}/F_{exp}, whereas hitchhiking or deleterious allele selection tend to result in reduced values of F_{obs}/F_{exp}.

As might be expected if the QTL-linked loci were influenced by selection, for both the N'Dama and Djakoré the lowest mean value of F_{obs}/F_{exp} is that for background loci, with F_{obs}/F_{exp} values for the three QTL regions all higher. However, this pattern is not seen in Somba, for which mean F_{obs}/F_{exp} for the BTA7 loci is lower than for the background loci. Additionally, the background loci show the lowest mean F_{obs}/F_{exp} in both European breeds, even though these have not been selected for trypanotolerance. Within each QTL region and amongst the background loci there is very high inter-locus variation in F_{obs}/F_{exp} (**Figure 3.6**). This means that there is no significant difference (as measured by t tests and Mann-Whitney tests) between F_{obs}/F_{exp} values for background loci and QTL loci in any breed.

F_{obs}/F_{exp} values for the purebred West African taurine breeds can be compared with those of the hybrid breeds derived from them. For N'Dama, mean F_{obs}/F_{exp} across all 33 loci is 0.8 ± 0.18 while that for Djakoré is 0.91 ± 0.30. A paired t test (comparing values between breeds for each locus in turn) shows that F_{obs}/F_{exp} values are significantly lower in N'Dama ($p<0.01$). For Somba and the derived hybrid Somba, the position is reversed. Mean

F_{obs}/F_{exp} across all loci in Somba is 0.92 ± 0.22, while the mean value for hybrid Somba is lower at 0.79 ± 0.25. A paired t test shows that the difference is significant ($p<0.01$).

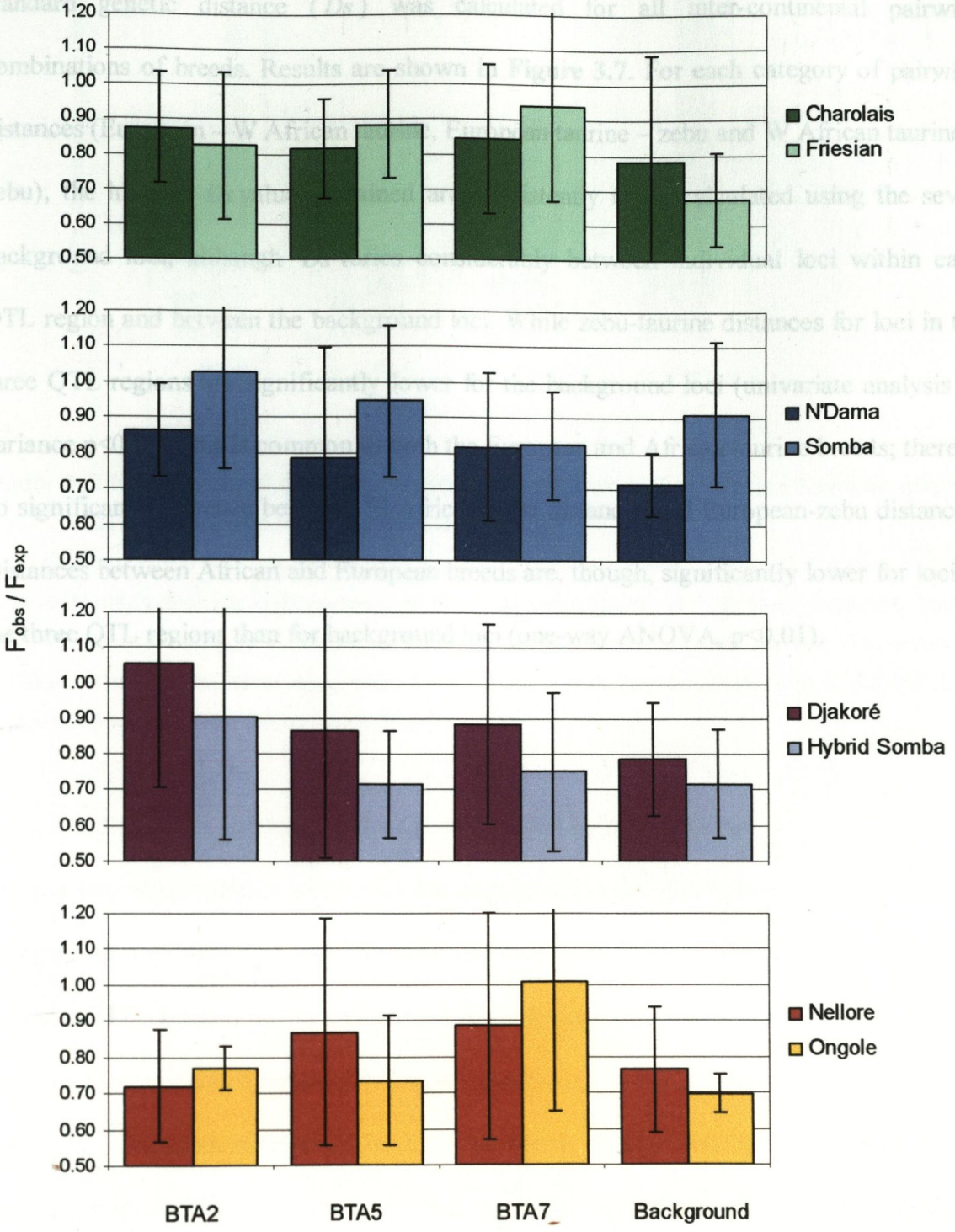

Figure 3.6: Departure from MDE for three candidate trypanotolerance QTL regions and background loci in trypanotolerant and susceptible cattle breeds

Fobs/Fexp results are averages of results for 9 BTA2 loci, 4 BTA5 loci, 10 BTA7 loci and 7 background loci

3.3.5 Inter-breed genetic distance

For the six purebred breeds (i.e. all except the Djakoré and the hybrid Somba), Nei's standard genetic distance (Ds) was calculated for all inter-continental pairwise combinations of breeds. Results are shown in **Figure 3.7**. For each category of pairwise distances (European – W African taurine, European taurine – zebu and W African taurine – zebu), the highest Ds values obtained are consistently those calculated using the seven background loci, although Ds varies considerably between individual loci within each QTL region and between the background loci. While zebu-taurine distances for loci in the three QTL regions are significantly lower for the background loci (univariate analysis of variance $p<0.01$), this is common to both the European and African taurine breeds; there is no significant difference between W-African-zebu distances and European-zebu distances. Distances between African and European breeds are, though, significantly lower for loci in the three QTL regions than for background loci (one-way ANOVA, $p<0.01$).

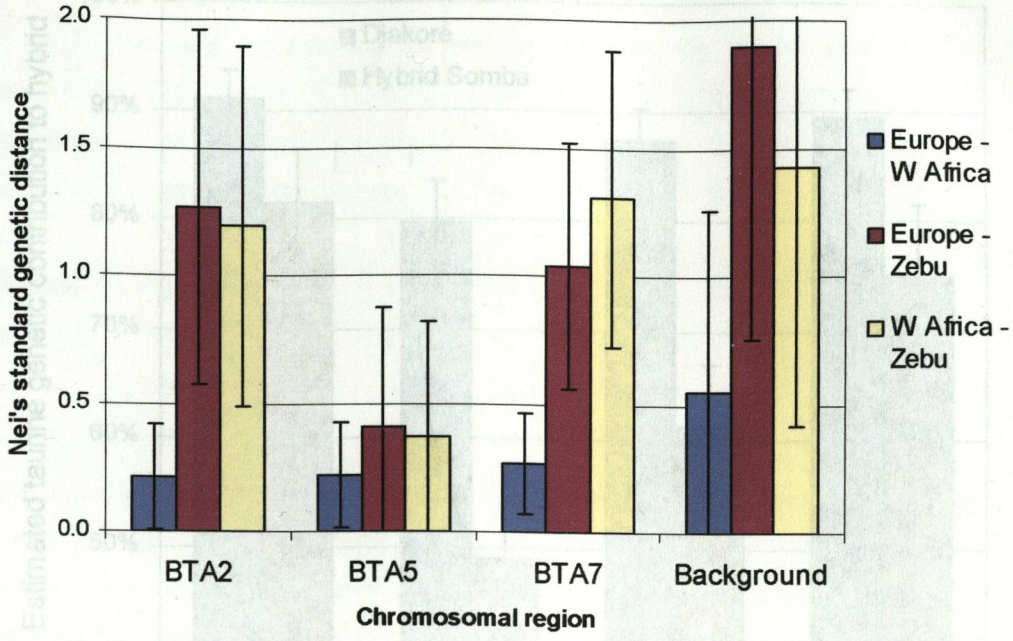

Figure 3.7: Nei's standard genetic distance calculated for loci from three trypanotolerance QTL regions and background loci

Each distance is the average of the four pairwise distances between purebred breeds – e.g. the Europe-W Africa distance is the average of the Charolais-N'Dama, H-F-N'Dama, Charolais-Somba and H-F -Somba distances (distances for West African hybrid populations are not included). Standard errors shown were calculated as the square root of the sum of squares of the individual bootstrapped errors obtained for each distance value.

3.3.6 Admixture estimates for West African hybrid breeds

For the two West African hybrid breeds studied (Djakoré and hybrid Somba), separate estimates of the relative zebu and taurine genetic contributions to the breed were obtained for each QTL region and the background loci. Results are shown in **Figure 3.8**. For the Djakoré, the estimates of taurine (N'Dama) genetic input vary between 80% and 91%. For the hybrid Somba, the estimates of genetic input from pure Somba are lower, reflecting the greater historical gene flow coming from zebu cattle, and vary much more (49% to 82%). In both breeds, the lowest estimate of taurine genetic input is obtained for loci in the BTA2 QTL region. **Figure 3.7** shows that genetic differentiation between zebu and taurine cattle is lowest for loci in this region, and this appears a likely cause of the low estimates of taurine genetic input obtained.

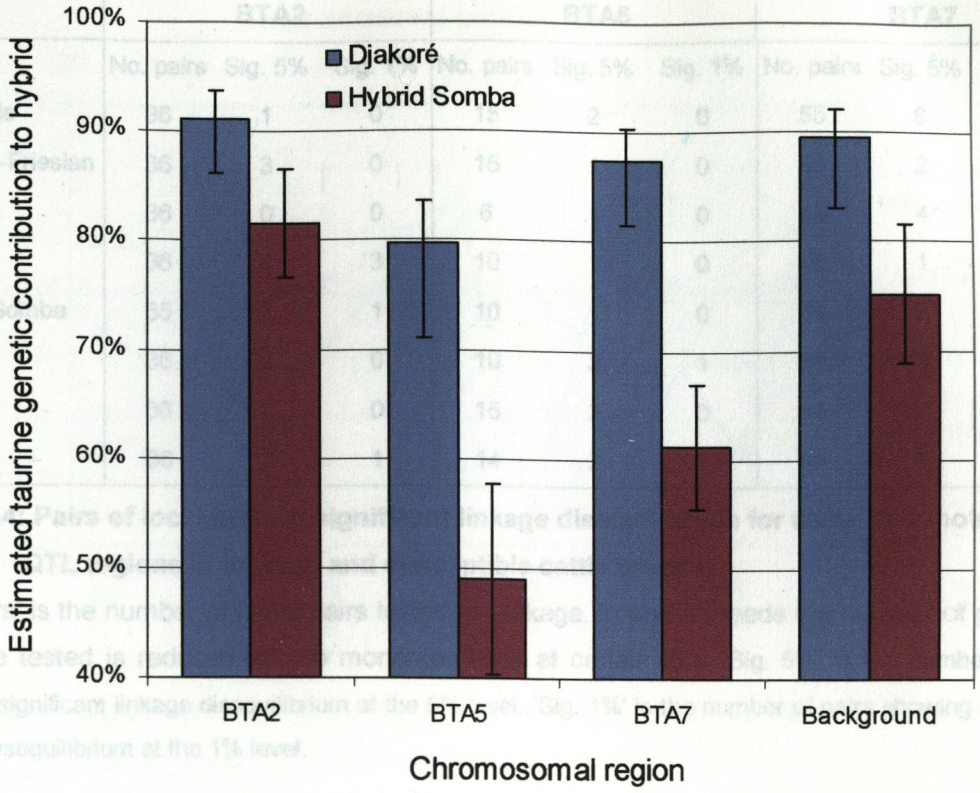

Figure 3.8: Degree of taurine genetic input into two West African hybrid populations at three candidate trypanotolerance QTL regions and background loci

3.3.7 Linkage disequilibrium across QTL regions

The proportion of locus pairs in significant linkage disequilibrium in each QTL region does not appear to differ between African and other breeds (**Table 3.4**). Also, parametric (one-way ANOVA) and non-parametric (Kruskal-Wallis) tests find that there are no significant differences in linkage probabilities between breeds. Interestingly the hybrid breeds do not show increased linkage disequilibrium relative to their presumed ancestral breeds.

	BTA2			BTA5			BTA7		
Breed	No. pairs	Sig. 5%	Sig. 1%	No. pairs	Sig. 5%	Sig. 1%	No. pairs	Sig. 5%	Sig. 1%
Charolais	36	1	0	15	2	0	55	6	2
Holstein-Friesian	36	3	0	15	1	0	55	2	0
N'Dama	36	0	0	6	1	0	45	4	1
Somba	36	4	3	10	1	0	45	1	1
Hybrid Somba	35	2	1	10	1	0	54	5	1
Djakoré	36	2	0	10	3	1	55	6	3
Nellore	36	2	0	15	2	0	54	7	2
Ongole	36	3	1	14	2	1	55	3	1

Table 3.4: Pairs of loci showing significant linkage disequilibrium for three trypanotolerance QTL regions in tolerant and susceptible cattle breeds

'No. pairs' is the number of locus pairs tested for linkage. In some breeds the number of pairs that could be tested is reduced due to monomorphism at certain loci. 'Sig. 5%' is the number of pairs showing significant linkage disequilibrium at the 5% level. 'Sig. 1%' is the number of pairs showing significant linkage disequilibrium at the 1% level.

3.3.8 Intra-QTL analysis results for individual loci

As well as comparing results for QTL regions with those for background loci, it is possible to compare individual loci within each QTL to see if any striking trends emerge.

3.3.9 Gene diversity and selective neutrality for BTA2 loci

Whereas it might be hoped that gene diversity in Somba and N'Dama would decline towards the selected locus, plotting gene diversity values for all four purebred taurine breeds (**Figure 3.9**) shows that there is no clear trend moving from locus to locus. It is even very difficult to discern any individual loci for which Somba and N'Dama both show reduced diversity while the European breeds do not.

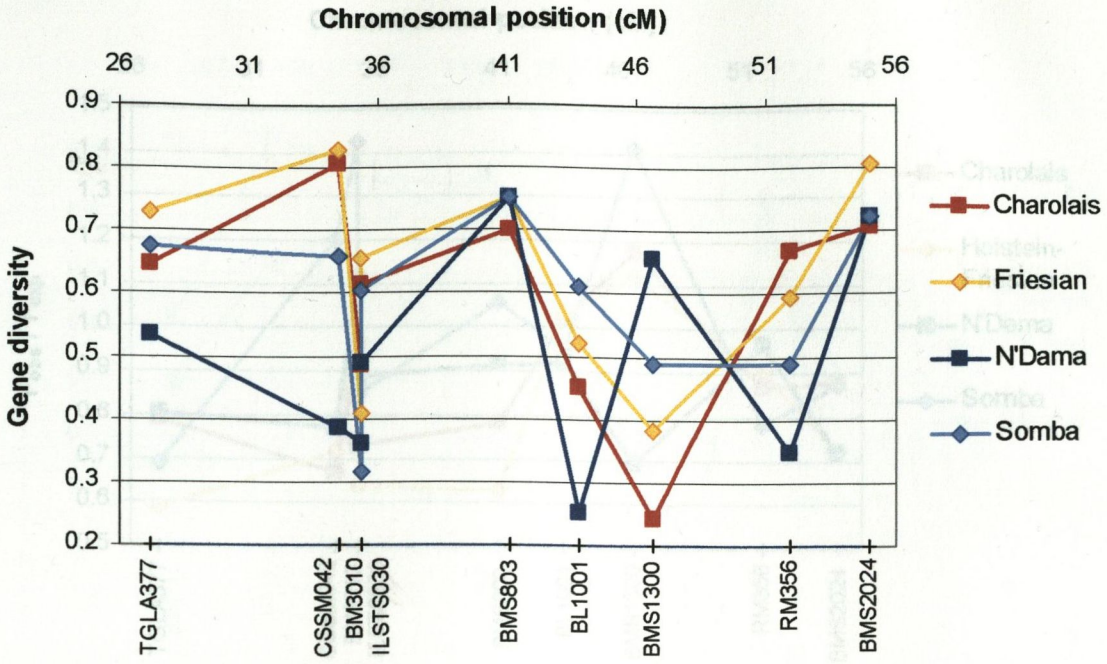

Figure 3.9: Gene diversity in 4 *B. taurus* breeds for 9 microsatellite loci from a trypanotolerance QTL region on BTA2

Positions of loci are taken from Kappes *et al.* (1997)

The picture for deviation from MDE (**Figure 3.10**) is equally chaotic. Values of F_{obs}/F_{exp} for Somba and N'Dama do not vary consistently and do not differ appreciably from those for the European breeds.

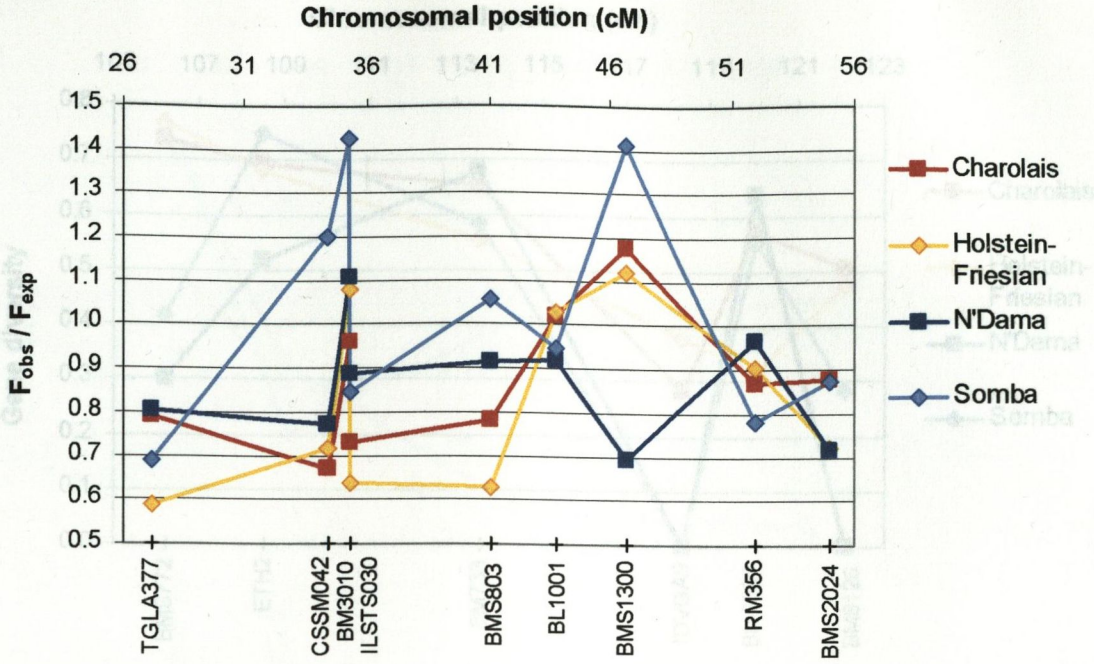

Figure 3.10: Departure from MDE in 4 *B. taurus* breeds for 9 microsatellite loci from a trypanotolerance QTL region on BTA2

Positions of loci are taken from Kappes *et al.* (1997)

3.3.10 Gene diversity and selective neutrality for BTA5 loci

For BTA5, locus IDVGA9 is monomorphic in the two African breeds, but not in the European breeds (**Figure 3.11**). This could be due to fixation at IDVGA9 due to selection at a linked gene. However, it is equally possible that the locus was never polymorphic in African taurine cattle. Supporting this is the fact that there are only two distinct alleles in the European breeds. Additionally, the flanking loci BM733 and BMS597 do not show low gene diversity in either Somba or N'Dama. For the loci at either end of the region (BMS772 and BM8126) there is a possibility that the reduced diversity seen in both West African breeds relative to the European breeds might indicate selection at a nearby gene. However, it would be unwise to infer this from single locus results.

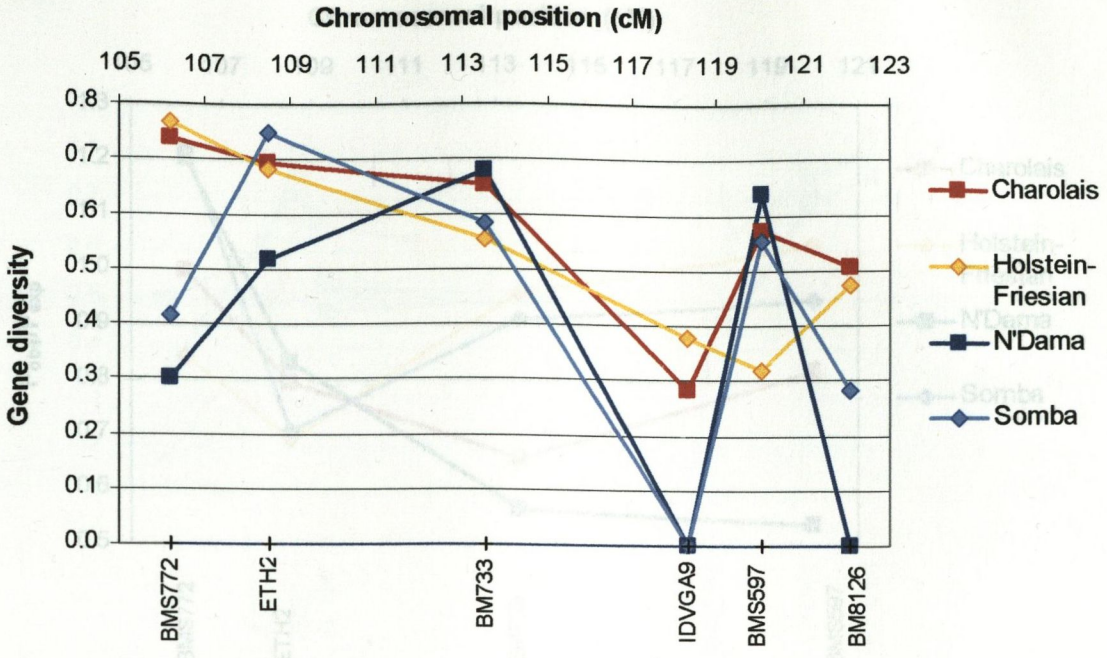

Figure 3.11 Gene diversity in 4 *B. taurus* breeds for 6 microsatellite loci from a trypanotolerance QTL region on BTA5

Positions of loci are taken from Kappes *et al.* (1997)

Only four loci were polymorphic in all four pure taurine breeds. Plotting values of F_{obs}/F_{exp} for these loci does not reveal any obvious differences between the African and European breeds (**Figure 3.12**), except possibly at locus BMS772, for which both African breeds show more skewed allele distributions than the European breeds. Again, inferring selection from a single locus result is ill-advised, though the gene diversity result for this locus referred to above, and the fact that the QTL is thought to be close to locus ETH2 (**Section 3.1.1b**), the locus nearest BMS772, suggests that further investigation of loci near BMS772 merits consideration.

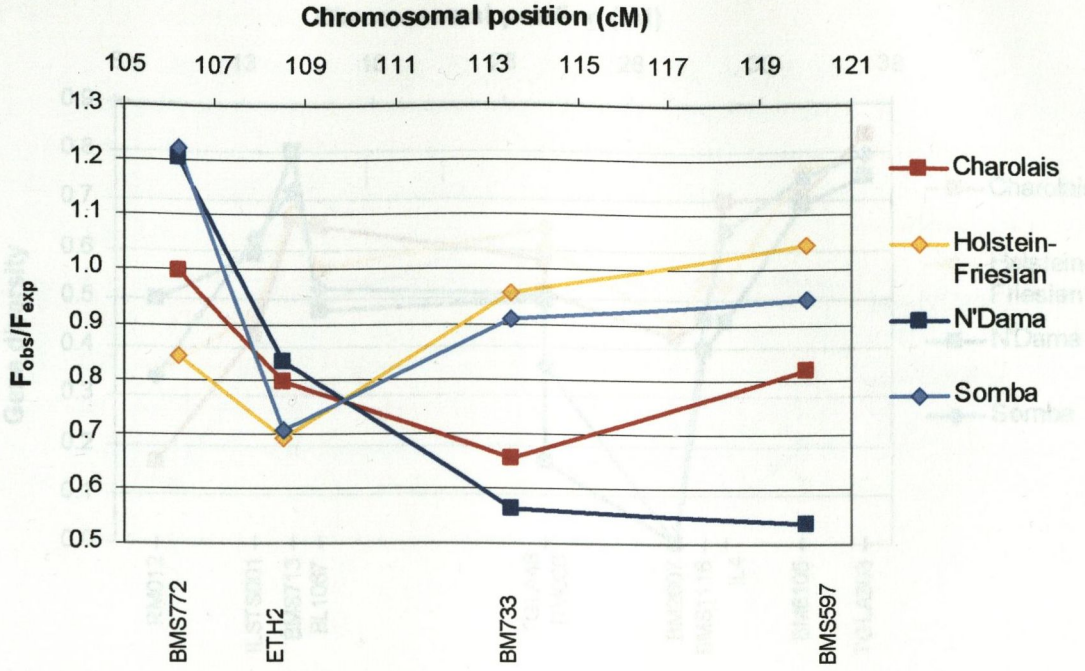

Figure 3.12: Departure from MDE in 4 *B. taurus* breeds for 4 microsatellite loci from a trypanotolerance QTL region on BTA5

Positions of loci are taken from Kappes *et al.* (1997)

3.3.11 Gene diversity and selective neutrality for BTA7 loci

A plot of gene diversity values for loci from the BTA7 QTL region (**Figure 3.13**) shows very low diversity at locus RM006, and monomorphism at the adjacent locus BM2607 in African breeds. European breeds do not show this pattern. For the remaining nine loci, though, there is no clear difference between gene diversity in African and European breeds.

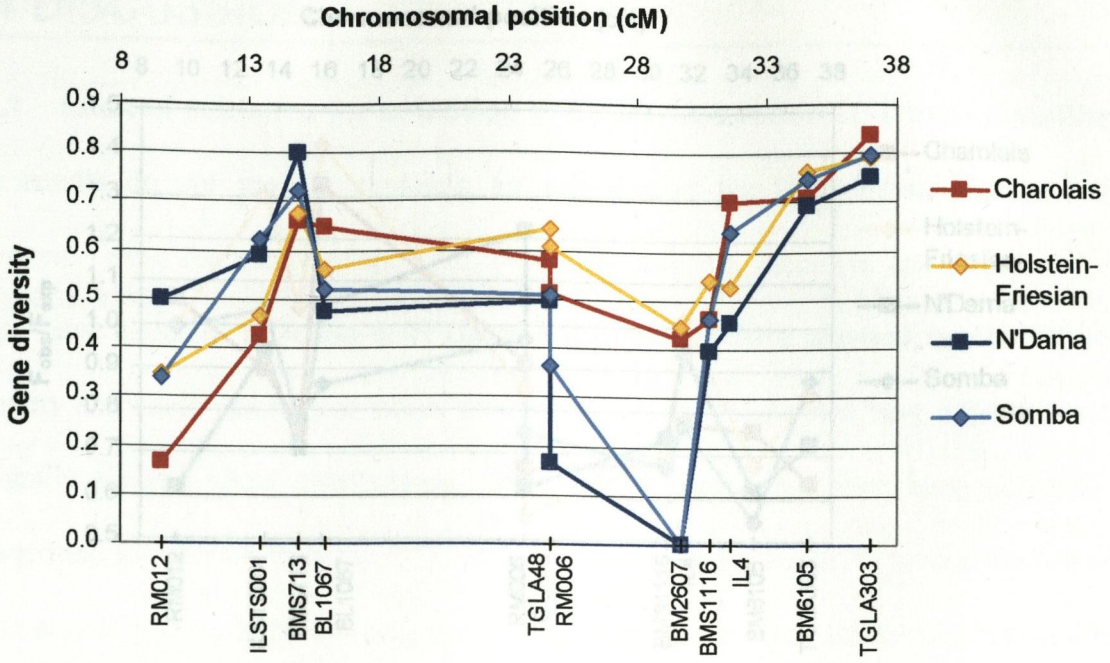

Figure 3.13: Gene diversity in 4 *B. taurus* breeds for 11 microsatellite loci from a trypanotolerance QTL region on BTA7

Positions of loci are taken from Gu *et al.* (2000). Locus TGLA48 is not included in the map of Gu *et al.*, and its position (fully linked to RM006) is taken from Hanotte *et al.* (in preparation).

Comparing departure from MDE for individual loci (BM2607 excluded due to monomorphism in African breeds) reveals that the F_{obs}/F_{exp} values in African breeds exceed those for European breeds at locus TGLA48 only (**Figure 3.14**). Allele distributions for RM006, which shows reduced gene diversity in N'Dama and Somba, are not markedly skewed - if anything the opposite.

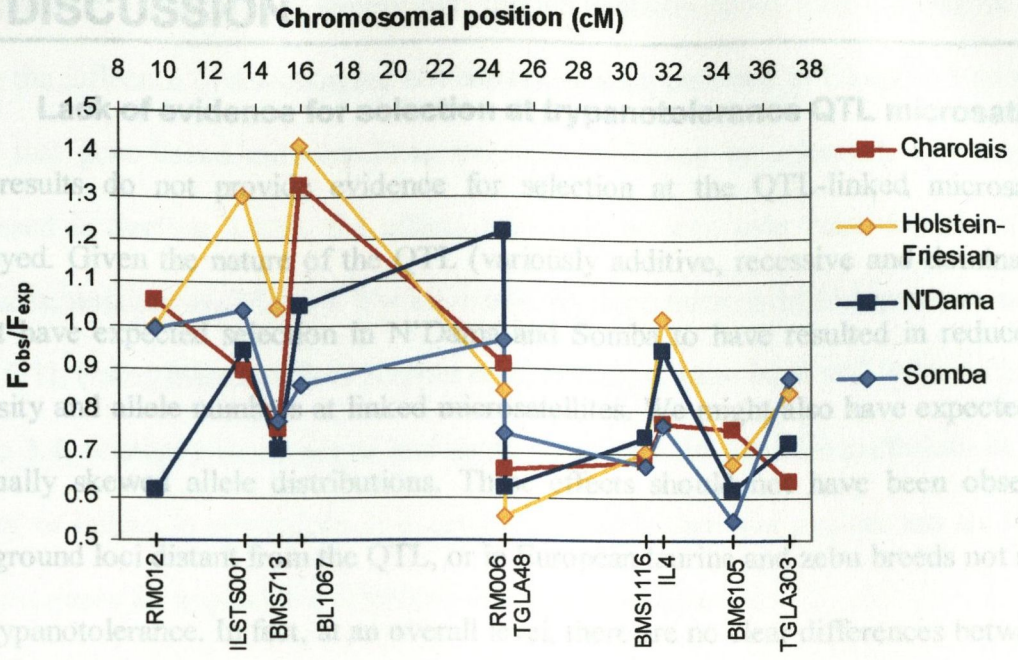

Figure 3.14: Departure from MDE in 4 *B. taurus* breeds for 10 microsatellite loci from a trypanotolerance QTL region on BTA7

Positions of loci are taken from Gu *et al.* (2000). Locus TGLA48 is not included in the map of Gu *et al.*, and its position (fully linked to RM006) is taken from Hanotte *et al.* (in preparation).

3.4 DISCUSSION

3.4.1 Lack of evidence for selection at trypanotolerance QTL microsatellites

The results do not provide evidence for selection at the QTL-linked microsatellites surveyed. Given the nature of the QTL (variously additive, recessive and dominant), we might have expected selection in N'Dama and Somba to have resulted in reduced gene diversity and allele numbers at linked microsatellites. We might also have expected to see unusually skewed allele distributions. These effects should not have been observed at background loci distant from the QTL, or in European taurine and zebu breeds not selected for trypanotolerance. In fact, at an overall level, there are no clear differences between loci linked to the three QTL and the background loci in African breeds that are not also seen in European breeds. This applies to all of the parameters studied: gene diversity, number of alleles per locus, deviations from MDE and genetic distance from zebu breeds.

Looking at a finer scale, within each QTL region, there are no obvious patterns of declining diversity or increasing departure from MDE towards an extreme value which might represent the closest locus to a selected gene. Instead, there is considerable variation from one locus to the next with little sign that diversity or allele distribution skewness at a particular locus is correlated with that of adjacent loci. Additionally the different variables do not appear to show good correlation for each locus. Individual locus plots of other variables (number of alleles pre locus, zebu-taurine genetic distance, and admixture estimates for hybrid breeds) are not shown here, but results are qualitatively similar to those for gene diversity and deviation from MDE shown in that there is considerable stochasticity and little evidence of clear trends in African cattle. There are several factors which may explain the lack of selective evidence.

One possible explanation is simply that the assumptions about microsatellite behaviour under the influence of selection are not correct. Results obtained in **Chapter 2** do seem to show that gene-linked microsatellites are indeed affected by selection. However, as is discussed in **Section 2.4.10**, the effects may only be seen over relatively short genetic distances, possibly as little as a few kilobases. Average microsatellite separation across all three QTL (using distances from Kappes *et al.* (1997)) is quite large at 3.2cM. As is seen in **Table 3.4**, relatively few pairs of loci show significant linkage disequilibrium in African breeds, or indeed in any breeds. It is certainly possible then that loci are too far from any selected genes for appreciable effects of selection to be detected.

A second possibility is that the QTL investigated are not genuinely associated with trypanotolerance. This seems less likely, particularly given the very high Lod scores for combined trypanotolerance traits seen for the BTA2 QTL (**Section 3.1.1a**) and the BTA7 QTL (**Section 3.1.1c**). One uncertainty, though, is that trypanotolerance in N'Dama and in Somba is due to the same genes. If different genes were involved in Somba, then we would not expect microsatellites at the N'Dama QTL identified from the ILRI trypanotolerance mapping pedigree to show selection in Somba.

3.4.2 Reduced diversity in N'Dama

Despite the lack of evidence for selection discussed, N'Dama and Somba cattle show consistently low diversity across all 26 QTL loci typed. Mean gene diversity is far lower in N'Dama than in any other breed, with that in Somba second lowest (**Table 3.5**).

The result for N'Dama is markedly lower than that of 0.503 ± 0.053 obtained by MacHugh (1996) for the same population using 20 different microsatellite loci. By contrast, the gene diversity figures reported by MacHugh (1996) for the same Charolais (0.549 ± 0.055) and

Friesian (0.545 ± 0.052) populations are lower than those reported here. This might suggest that the QTL-linked loci do indeed show reduced diversity in N'Dama. Possibly this was not detected in this study because the seven background loci were in some way non-representative.

Population	Gene diversity
N'Dama	0.460 ± 0.046
Somba	0.524 ± 0.041
Nellore	0.532 ± 0.047
Ongole	0.554 ± 0.041
Charolais	0.573 ± 0.033
Friesian	0.586 ± 0.030

Table 3.5: Mean gene diversity values for 26 microsatellite loci across three QTL regions in six purebred cattle breeds

3.4.3 Deviation from MDE in admixed African breeds

The two African hybrid breeds studied exhibit different levels of zebu genetic introgression, with the Djakoré showing a much lower level than the hybrid Somba (**Section 3.3.6**). Over all loci, the Djakoré are estimated to have 10.8% ± 1.6% zebu genes, while the hybrid Somba have 28.3% ± 1.4% zebu genes. As shown by simulations of hybrid populations (**Section 2.3.10**), artificially constructed hybrid cattle populations with up to 20% zebu genes have allele distributions which appear to deviate from MDE in the direction of excess homozygosity. For artificial hybrid populations with 20 to 50% zebu genes, deviation from MDE is in the direction of heterozygosity excess. The real data from the Djakoré and hybrid Somba agree with the simulation results. **Figure 3.6** shows that for all three QTL and background loci, Djakoré tend towards homozygosity excess when compared with N'Dama, whereas the hybrid Somba tend towards homozygosity deficiency when compared with pure Somba.

CHAPTER 4

DEVELOPMENT AND MAPPING OF SNP LOCI WITHIN A CANDIDATE REGION FOR TRYPANOTOLERANCE ON BOVINE CHROMOSOME 7

4.1 INTRODUCTION

4.1.1 Requirement for new polymorphic markers to refine QTL positions

In the trypanotolerance genome scan undertaken at ILRI, 20 microsatellites were typed on chromosome BTA7. The microsatellites span a distance of approximately 96cM, with an average inter-locus distance of 4.8cM. However, the limited number of recombination events in the pedigree means that the microsatellites map to only 12 distinct locations, with an average separation of 8cM. To refine the position of the QTL using pedigree analysis it would be necessary to create an F3 generation to increase the number of recombinations in the pedigree. It would also be necessary to increase the number of markers genotyped, particularly within the candidate QTL region in the centromeric part of the chromosome. At present, the number of loci mapped to the QTL region is limited, hence an aim of the work described here was to identify new loci. Clearly any new markers typed must be polymorphic to detect associations between phenotypic traits and alleles. Ideally, new markers should be fixed for different alleles in zebu and taurine as such markers are more likely to prove informative in QTL mapping. Such markers will also be useful in studies of field populations from the Sahelian zebu / taurine hybrid zone. As was seen in **Chapter 3**, the inherent stochasticity of microsatellite loci limits their utility in dissecting variation in levels of population admixture at the intra-chromosomal scale. The development of a panel of linked, biallelic markers showing perfect discrimination between zebu and taurine cattle would greatly facilitate this type of analysis.

It is also desirable to add new genes to the BTA7 map to enable comparison of the bovine map with the considerably more advanced human genome map. Understanding the relationship between the two maps allows prediction of gene location in cattle based on location in human, which may assist in identifying candidate trypanotolerance genes. A

second aim of the work described in this chapter was therefore to add new type I loci to the increasingly dense physical map of BTA7.

4.1.2 Finding novel loci on BTA7 using a chromosome-fragment library

One approach to finding new loci within a defined chromosomal region is to use chromosome micro-dissection to prepare a library of DNA fragments from the defined region (Goldammer *et al.* 1996). Here, several copies of the target chromosome portion are obtained through repeated micro-dissection of prepared chromosomes with ultra-fine glass capillaries. The chromosome portions are then used as template for PCR amplification using degenerate primers to create a library of short DNA sequences (Goldammer *et al.* 1996). The sequences are ligated into plasmids and grown up in bacterial cell culture. Plasmids are harvested and the inserts are then sequenced. The method has potential problems, however. Use of three or four copies of template (about 50fg) for degenerate oligo PCR creates a high risk of contamination, so the final library may contain DNA sequences extraneous to the target genomic region (Goldammer *et al.* 1996). Also, the use of degenerate PCR primers can give rise to PCR artefacts. A final problem is that DNA sequences could come from anywhere within the chromosomal fragment, which is likely to be many megabases of DNA in size. A large number of sequences may therefore have to be screened to find loci within a precise region of interest.

4.1.3 Finding polymorphic loci on BTA7 through comparative genomics

Another way of obtaining new polymorphic loci within a target chromosome region is to search for polymorphism associated with genes.

According to the neutral theory of molecular evolution, the rate of evolution at a nucleotide site relates to the functional constraint on the site (Kimura 1983). This means that it should

be more profitable to search for polymorphism within introns than within exons. This has been confirmed by a study comparing intron and exon sequences in rats and mice which showed the nucleotide substitution rate in introns to be approximately equal to the substitution rate for synonymous sites of exons, and approximately three times the rate for non-synonymous sites of exons (Hughes and Yeager 1997; Hughes and Yeager 1998).

A problem is that intron positions are frequently unknown in cattle, as often only cDNA sequences are available. To locate a given intron in cattle, it is necessary to refer to another mammalian species for which the genomic sequence is available. Intron position is highly conserved in vertebrates. A study has found that, for 25 genes sequenced in human and cattle, 73 of 75 introns were present in both species, (K. H. Wolfe, personal communication). Aligning the cattle and human sequences thus reveals the position of the intron in the bovine cDNA sequence. Once the intron position has been identified in cattle, primers can be designed in the exon DNA either side of the intron to amplify across the intron.

If insufficient mapped genes are available in a specific bovine chromosomal region, it is now possible to exploit information from comparative genomics to predict which genes may be present. Primers can then be designed using gene sequences from other mammals (typically human or mouse) to amplify the orthologous gene sequence in cattle. The bovine sequence amplified can now be mapped relatively easily to verify if it is in the predicted location. Notably, this has only become possible through recent advances in bovine genome mapping.

4.1.4 Bovine genome mapping

Bovine genome mapping has lagged considerably behind mapping in human and mouse, which are the subjects of major genome sequencing programs. The main mapping method employed until recently, genetic linkage mapping, does not readily permit comparative mapping between species as it utilises chiefly type II loci selected for polymorphism, not coding sequences. However, recent advances, first in cross-species 'chromosome painting', and latterly in radiation hybrid mapping have enabled comparison of genome organisation in cattle and human or mouse.

4.1.5 Genetic linkage mapping

Linkage maps rely for their creation on coinheritance of alleles at different loci in a pedigree of individuals. The further apart two loci are, the more likely it is that a recombination event will occur between them. Frequent coinheritance of alleles at two distinct loci therefore implies that the loci are closely linked.

Over the last decade, a number of bovine whole-genome genetic linkage maps have been published. The initial two maps were published in 1994; the first contained 202 loci covering 2,513cM, or approximately 90% of the genome (Barendse *et al.* 1994) and the second contained 313 loci covering 2,464cM (Bishop *et al.* 1994). Continued development of new bovine genetic markers has allowed construction of more detailed maps. Two maps were published in 1997, the first containing 746 loci with an average spacing of 5.3cM (Barendse *et al.* 1997) and the second 1,250 loci with an average spacing of 2.5cM (Kappes *et al.* 1997). The map of Kappes *et al.* includes 41 loci from BTA7 spanning a length of 137.3 cM (an average spacing of ~3.3cM). By combining data from eight different bovine pedigrees, Gu *et al.* (2000) produced a comprehensive BTA7 linkage map including 54 loci spanning a length of 158.6cM (average spacing ~2.9cM).

Although extensively used, linkage mapping has significant limitations. One problem is that map resolution is limited by the number of recombination events in the mapping pedigree. Resolution can thus only be increased by expanding the pedigree, which entails breeding further generations. This is expensive and requires years of waiting.

A second problem is that only polymorphic loci can be mapped, hence researchers seeking new polymorphic markers for mapping have found it most profitable to concentrate on microsatellites. These are highly polymorphic, and relatively easy to find using DNA libraries enriched for dinucleotide repeats. Of the 54 loci in the comprehensive BTA7 linkage map, 48 are microsatellites (Gu *et al.* 2000). While microsatellites are often conserved between relatively closely related species such as cattle, sheep and goat, they are not conserved in distantly related species. Linkage maps based predominantly on microsatellites are therefore of limited use for comparing gene organisation in cattle and humans or mice.

One strategy for linkage mapping of genes used an interspecific hybrid *Bos taurus* x *Bos gaurus* pedigree (Gao and Womack 1997b). The large evolutionary separation between the species means that there is a relatively high probability of polymorphism, even within coding sequence. Although the pedigree used was of limited resolution, Gao *et al.* were able to determine the order on BTA7 of six genes.

4.1.6 Comparative gene mapping

To predict which genes might be present in a particular bovine genomic region, it is necessary to understand the relationship between the bovine genome and the human genome. The other mammal whose genome is very well characterised is the mouse, and the

organisation of the bovine genome is more similar to that of the human than the murine genome. Comparisons of chromosomal assignments of genes in human, mouse and cattle show a higher degree of conservation between the bovine and human genomes than between the bovine and mouse genomes (Womack and Moll 1986). Recent data from the Mouse Genome Database show that genes from each human chromosome are distributed across an average of 2.4 bovine chromosomes (257 genes), while loci from each mouse chromosome are distributed across an average of 4.0 bovine chromosomes (237 genes) [homology data retrieved November 2000 from the Mouse Genome Database (MGD), URL: http://www.informatics.jax.org/]. The location of a gene in cattle can therefore be more accurately predicted using the location in human than that in mouse. In the case of BTA7, orthologous genes to those assigned on the chromosome are located mostly on human chromosome 5 (HSA5) and human chromosome 19 (HSA19) (Eggen and Fries 1995; O'Brien *et al.* 1993; Zhang and Womack 1992).

4.1.7 Cross-species chromosome painting (Zoo-FISH)

A technique that has proved particularly informative in comparative genome mapping between cattle and other human is cross-species chromosome painting (also known as Zoo-FISH). This involves probing a bovine chromosome spread using a library of sequences from a specific human chromosome to detect regions of extensive homology (Chowdhary *et al.* 1996; Solinas-Toldo *et al.* 1995). For BTA7, sequences from HSA19 hybridise to the centromeric region (7q12-15 or 7q12-21), while sequences from HSA5 hybridise to the distal region (7q21-28 or 7q22-28) (Chowdhary *et al.* 1996; Solinas-Toldo *et al.* 1995). Radiation hybrid mapping, however, is necessary to detect rearrangements occurring within these broad regions of conserved synteny.

4.1.8 Radiation hybrid mapping

The application of radiation hybrid mapping in cattle has transformed comparative mapping. Radiation hybrid mapping makes use of the ability of high-energy radiation to physically break chromosomes. The closer two loci are on a chromosome, the less likely that a break will occur between them on irradiation.

Radiation hybrid mapping was developed from the early work of Goss and Harris, who demonstrated that gamma radiation could induce physical breaks between loci on the human X chromosome (Goss and Harris 1975). Cox *et al.* demonstrated that the fragments of human chromosome obtained through irradiation of a human-hamster somatic cell hybrid containing only a single human chromosome (HSA21) could be used to construct a map of the chromosome (Cox *et al.* 1990). Walter *et al.* adapted the method used by of Cox *et al.* to create whole-genome radiation hybrid mapping panels that could be used to map any chromosome (**Figure 4.1**) (Walter *et al.* 1994). They demonstrated that a human fibroblast culture (which contains a full complement of human chromosomes) could be irradiated using X-ray radiation, fragmenting the human chromosomes. As in the method of Cox *et al.*, it is necessary to rescue the irradiated cells from dying by fusing them with Chinese hamster cells. After fusion, some of the human chromosome fragments come to be present in a background of hamster chromosome. Some fragments remain independent, and may survive in the cell line if they have the necessary elements for correct segregation at cell division. The hamster cells used lack the thymidine kinase (TK) gene, so selection for the presence of the human TK gene eliminates hamster cells that contain no human chromosomal fragments.

To create a mapping panel, cell lines are cultured from different rescued cells, each containing different fragments of the original human chromosomes. For mapping, each line

is scored for presence or absence of human loci. Loci that are frequently present together in the same cell line (co-retained) are inferred to be physically close on the chromosome, with loci showing lower co-retention further apart.

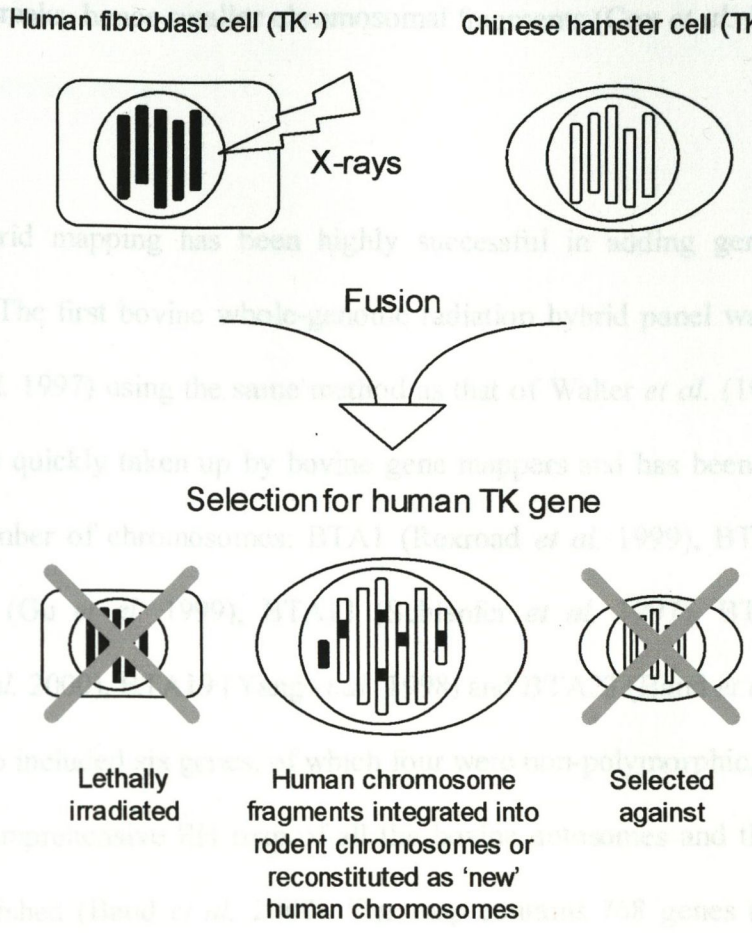

Figure 4.1: Fusion of irradiated human cells with hamster cells to create a radiation hybrid panel for gene mapping

Adapted from Walter *et al.* (1994)

Whole-genome radiation hybrid mapping has become a widely used technique, with panels available for a number of vertebrate species. In human, a map of 30,000 genes has been produced using two whole-genome RH panels (Deloukas *et al.* 1998).

Radiation hybrid mapping has a number of key advantages over linkage mapping. One major advantage is that radiation hybrid mapping does not require loci to be polymorphic,

so genes showing no allelic variation may be mapped. This greatly facilitates comparative genomics, as genes mapped in human or mice can be mapped in cattle. A second advantage is that map resolution can be controlled by selecting the dose of radiation used to create the panel. A higher dosage of radiation results in a greater number of chromosome breaks, hence smaller chromosomal fragments (Cox *et al.* 1990; Walter *et al.* 1994).

Radiation hybrid mapping has been highly successful in adding genes to the bovine genome map. The first bovine whole-genome radiation hybrid panel was created in 1997 (Womack *et al.* 1997) using the same method as that of Walter *et al.* (1994) (**Figure 4.1**). The panel was quickly taken up by bovine gene mappers and has been used to construct maps of a number of chromosomes: BTA1 (Rexroad *et al.* 1999), BTA5 (Ozawa *et al.* 2000), BTA7 (Gu *et al.* 1999), BTA13 (Schlapfer *et al.* 1997), BTA15 and BTA29 (Amarante *et al.* 2000), BTA19 (Yang *et al.* 1998) and BTA23 (Band *et al.* 1998). The first BTA7 RH map included six genes, of which four were non-polymorphic, and 28 other loci. Recently, a comprehensive RH map of all the bovine autosomes and the X chromosome has been published (Band *et al.* 2000). The map contains 768 genes (of which 358 are expressed sequence tag, or 'EST' sequences) and 319 microsatellites. Loci mapped on BTA7 in this map include 39 genes (of which 10 are EST sequences) and 14 microsatellites (Band *et al.* 2000). The development of a second, high-resolution whole-genome panel (Rexroad *et al.* 2000) holds the promise of more detailed maps.

4.1.9 Predictive gene mapping

The current understanding of comparative genomics enables confident prediction of gene location in cattle based on location in human. A recent study confirmed that the bovine orthologues of four genes from the p arm of HSA19 all map to BTA7 (Gu *et al.* 1997).

Radiation hybrid mapping confirmed that the genes map, as predicted, to the centromeric region (Gu *et al.* 1999). With 30,000 genes mapped using radiation hybrid panels in human (Deloukas *et al.* 1998), and the ongoing assembly of the human genome sequence (see e.g. http://www.ensembl.org/), there is a huge number of loci whose locations may be confidently predicted in cattle, allowing researchers to build up highly detailed maps of specific chromosome regions.

4.2 MATERIALS AND METHODS

4.2.1 Development of novel BTA7 loci using a chromosome-fragment library

For generating novel polymorphic markers on BTA7, a chromosome-fragment specific library of the candidate trypanotolerance region of BTA7q14-q21 was provided by Goldammer *et al.* (1996) This library was created by micro-dissection of chromosome BTA7, followed by PCR amplification using the degenerate primer 5'-CCGACTCGAGNNNNNNATGTGG-3' and ligation of amplified fragments into the pBluescript-derived plasmid 'PCR script AMP' (Stratagene) (Goldammer *et al.* 1996).

4.2.2 Transformation of *E. coli* with BTA7-derived library

The BTA7 library was used to transform Epicurian Coli XL10-Gold Ultracompetent Cells (Stratagene) as per the manufacturer's instructions:

1. Epicurian Coli XL10-Gold Ultracompetent cells were thawed on ice.
2. Cells were mixed by hand, and 40µl was aliquotted into two prechilled 15-ml polypropylene tubes
3. 1.6µl of the beta-Mercaptoethanol mix provided was added to each aliquot of cells.
4. The contents of the tube were swirled gently and incubated on ice for 10 minutes, swirling gently every 2 minutes.
5. 1µl of the BTA7 library was added to one of the cell aliquots before mixing gently. As a control, 1µl of the pUC18 control plasmid provided (diluted 1:10 in sterile water) was added to the other cell aliquot.
6. The tubes were incubated on ice for 30 minutes.
7. NZY^+ broth (1% casein digest (NZ amine), 0.5% NaCl, 0.5% yeast extract, 1% Bacto casamino acids, 0.2% MgSO4.7H2O, 0.5% NaCl.) was preheated in a 42°C water bath for use in step 10.

8. Cell aliquot tubes were heat-pulsed in a 42°C water bath for 30 seconds.
9. The tubes were incubated on ice for 2 minutes.
10. 0.45 ml of preheated (42°C) NZY^+ broth was added to each tube before incubation at 37°C for 1 hour with shaking at 250 rpm.
11. 0.4 ml of NZY^+ broth was added to each tube, and mixed. 100µl of transformed cells were plated onto LB-agar plates containing 100mg/ml ampicillin for selection for the plasmid and 40mg/ml X-Gal and 400µg/ml IPTG for blue / white selection for plasmid inserts.
12. Plates were incubated overnight at 37°C

White colonies were replated onto fresh LB agar Amp X-Gal IPTG plates by touching a sterile toothpick onto the colony and then onto the new plate. The rest of the colony was prepared for PCR. The colony was resuspended in 35µl lysis buffer (0.1% Triton-X, 10mM Tris pH 8.0), incubated at 95°C for 10 minutes and on ice for two minutes and spun for two minutes in a microfuge. 2µl of supernatant was used as template for PCR using the M13 forward and reverse primers incorporated either side of the PCR script Amp cloning site. Total reaction volume was 25µl. Other reagents were as previously described.

To determine the size of the inserts, 1µl of each PCR reaction was run out on a 1% agarose gel. Plasmids bearing the largest inserts were selected for plasmid preps.

4.2.3 Plasmid preps

1. White colonies were picked using a sterile toothpick and placed in10 ml L-Broth with ampicillin at 100mg/ml in sterile tubes.
2. Tubes were incubated overnight at 37°C whilst shaking at 225 rpm.
3. Cells were spun at 4500 rpm for 10 minutes, and the supernatant poured off.
4. Cells were resuspended in 60µl TES (50mM Tris pH 8.0, 25% sucrose, 2mM EDTA) and transferred to an eppendorf tube.
5. 20µl lysozyme (40mg/ml in 250mM Tris pH 8.0) was added.
6. Tubes were incubated on ice for 10 minutes
7. 550µl M-STET (5% Triton X-100, 50mM EDTA pH8.0, 50mM EDTA pH 8.0, 5% sucrose) was added.
8. Tubes were placed in a boiling water bath with their lids open for 90 seconds.
9. Tubes were spun for 30 minutes in a microfuge at full speed, and the resulting pellet removed.
10. 10µl heat-treated RNAse (10mg/ml) was added before incubation at 37°C for 30 minutes.
11. 500µl of phenol was added, and the tube vortexed thoroughly.
12. Tubes were spun for 5 minutes, and the upper phase was carefully transferred to a clean eppendorf.
13. 0.6 volumes of isopropanol were added and the tubes mixed and incubated at room temperature for 15 minutes.
14. The tubes were spun in a microfuge at full speed for 10 minutes and the isopropanol layer was removed.
15. 0.75ml 70% ethanol was added and the tubes were left for 5 minutes at room temperature.
16. Tubes were spun in a microfuge for 5 minutes, and the 70% ethanol was poured off. Tubes were spun again briefly and residual ethanol removed using a pipette.
17. DNA pellets were air-dried for 20 minutes at room temperature
18. DNA pellets were resuspended in 50µl sterile water, vortexing well.

2 µl of each plasmid solution was run out on a 1% agarose gel to check purity and determine approximate concentration. Nine plasmids containing inserts of between approximately 450bp and 750bp were selected for sequencing. Concentration was

determined by measuring OD260 (optical density at 260nm). The plasmid templates were sent for commercial sequencing using standard sequencing primers incorporated into the plasmid.

Sequences were checked for similarity to existing entries in GenBank using the BLAST® (Basic Local Alignment Search Tool) tool (Altschul *et al.* 1997), and for presence of common repeat elements using the on-line Repeat Masker tool (Smit).

4.2.4 Development of novel BTA7 loci by comparative mapping with human

Genes from human chromosome 19 were identified as candidates for the development of novel single nucleotide polymorphism (SNP) markers within the candidate genomic region on BTA7. This region is known to show homology to the p arm of HSA19 (Gu *et al.* 1999). The criteria used in identifying candidate loci are described below.

Genes identified as candidates span the centromeric half of HSA19 p arm. The region is approximately 56 cR_{3000} long, and the total length of HSA19 is approximately 309 cR_{3000} (Deloukas *et al.* 1998). Non-genomic sequences were discarded, and only sequences containing defined introns were considered. Intron sequences were checked for presence of common human repeats using the on-line Repeat Masker tool (Smit). Introns consisting mostly of repeat sequence were discarded. Remaining introns were screened for PCR suitability according to their size in human. Introns selected as candidates varied between 414bp and 727bp in human, with a total amplicon length in human of about 520 bp to 1100bp. Human genomic sequences were aligned with orthologous bovine or murine mRNA / cDNA sequences obtained from GenBank using the BLAST® tool (Altschul *et al.* 1997). Human sequences lacking a bovine or murine orthologue were discarded. If a bovine sequence was available, this was used to design PCR primers to anneal to the exons

on either side of the putative intron. If no bovine sequence was available, conserved primers were designed that would amplify the selected intron in both human and mouse, hence presumably in cattle. Conservation of primers at the 3' end was emphasised to reduce risk of PCR failure due to mismatch. Where mismatches occurred between human and mouse in the prospective primer, the human base was chosen.

4.2.5 PCR amplification

PCR was carried out for novel intronic loci and four previously published STS (sequence tagged site) loci (Gu *et al.* 1997) as described in **Section 2.2.6a**. Representative cattle DNA samples were used for PCR: one European taurine, one West African taurine, and one Indian zebu. For novel intronic loci designed from human sequence, a human DNA sample was used as a positive control. PCR yield and specificity were assessed by running products out on 1-2% agarose gels with a 100bp increment DNA ladder as a size standard.

For novel intronic loci where PCR using the initial forward and reverse primers was unsuccessful, additional PCR primers were designed to amplify the same intron. Further PCR using all possible combinations of forward and reverse primers was then attempted.

4.2.6 Sequencing

PCR products (obtained as described in **Section 2.2.6a**, but with a 50µl total volume) were cleaned for sequencing using a QIAquick PCR Purification Kit, as per the manufacturer's instructions (Qiagen, Valencia, CA). 50µl ddH_2O was used for the final elution of sequence template. Template concentration was determined from the OD260, and purity was verified by running out 2µl of the sample on a 1% agarose gel stained with ethidium bromide. Sequencing was carried out using a 'Big Dye' terminator cycle sequencing kit (Applied Biosystems, Foster City, CA). Reactions included: 3.2pmol of primer 50ng of

PCR product template, 6.0µl Big Dye reaction mixture, ddH_2O to a final volume of 20.0µl. Cycle sequencing was performed using a GeneAmp thermal cycler. An initial step of 95°C for 5 minutes was followed by 35 cycles of 95°C for 30sec, 50°C for 20sec, 60°C for 4 mins. Sequence reactions were precipitated by incubation at room temperature for 2 minutes after addition of 2µl 3M sodium acetate and 50µl absolute ethanol, followed by centrifugation for 20 minutes at 14,000 rpm. The pellets obtained were washed with 70% ethanol and air dried at 37°C. Samples were dissolved in 4µl redistilled formamide and 1µl of each was loaded onto a polyacrylamide gel for sequencing using an ABI sequencer.

4.2.7 Mapping

Prior to mapping using bovine-Chinese hamster somatic cell hybrid and radiation hybrid panels, loci were screened by PCR using bovine and Chinese hamster control DNA. Where bovine and hamster PCR products were indistinguishable on an agarose gel, new cattle-specific primers were designed using the bovine intron sequence.

4.2.8 Somatic cell hybrid mapping

Novel intron sequences were assigned to chromosomes by analysis of the PCR amplification pattern obtained from a panel of 31 hamster-bovine somatic cell hybrids derived from an original panel of 36 hybrids (Womack and Moll 1986) as previously described (Guerin *et al.* 1994). Bovine and Chinese hamster DNA samples were included as positive and negative controls. PCR products were run out on 2% or 3% agarose gels stained with ethidium bromide. Hybrids were scored for presence or absence of bovine specific amplification products. Analysis of results was carried out by Jan Johnson at Texas A & M University (J. S. Johnson, personal communication).

4.2.9 Radiation hybrid mapping

Two whole-genome radiation hybrid panels were used for mapping. Both were created from an Aberdeen Angus bull fibroblast line irradiated using gamma radiation from a ^{60}Co source. Irradiated cells were fused with a thymidine kinase (TK) deficient Chinese hamster fibroblast cell line. Fused cells were cultured in HAT medium to select for cells containing the bovine TK gene on chromosome BTA19. Surviving cells were cultured and DNA was extracted and tested for presence of bovine DNA (Womack *et al.* 1997). Different intensities of radiation were used in creating the two panels: 5,000 rad for panel BovR5 (Womack *et al.* 1997), and 12,000 rad for panel BovR12 (Rexroad *et al.* 2000). The final BovR5 panel contains DNA from 90 hybrid lines and the final BovR12 panel contains DNA from 180 hybrid lines. PCR amplification patterns were obtained for each locus for both panels. Contamination risk was minimised by using filter tips and performing all operations in a laminar flow hood. Bovine and Chinese hamster DNA samples were included as positive and negative controls. PCR products were run out on 3% agarose gels stained with ethidium bromide and hybrids were scored for presence or absence of bovine specific amplification products. PCR reactions for all hybrid-locus combinations were performed in duplicate as described in **Section 2.2.6** and results compared. Hybrid-locus combinations were retested if the initial two results conflicted.

Data was combined with previously obtained amplification data for loci typed in both BovR5 and BovR12 (Kang'a *et al.* 2000). Map locations of loci were determined using the software package RHmap version 3.0 (Lange *et al.* 1995; Lunetta *et al.* 1996). Assumptions made by RHmap are that radiation-induced breakage occurs randomly along the chromosome and that different chromosomal fragments are retained independently. Groups of linked loci were identified by two-point analysis of the BovR12 data using the program RH2pt with varying Lod score criteria from 1 to 15.

Likeliest locus orders were obtained for each linkage group using either data from both BovR5 and BovR12, or data from BovR12 alone. Locus order was determined using the program RHmaxlik, which calculates locus order likelihoods for each hybrid, given the observed chromosome breaks, and sums over all hybrids to obtain likelihoods for the entire panel (Boehnke *et al.* 1991). Locus orders were built using the 'stepwise' method of RHmaxlik whereby loci are added one by one to the list of saved n-locus partial locus orders. Maximum likelihoods are calculated for the new n+1-locus orders, and orders whose likelihoods are higher than a pre-defined threshold are retained for the next round of locus addition (Boehnke *et al.* 1991). The maximum likelihood threshold for order retention was set to 10^{-12} that of the best currently retained partial order.

For each linkage group, RHmaxlik was used to create a framework map consisting of loci added to the order at Lod >=3, such that the likelihood of the framework order is at least 1,000X the likelihood of the next likeliest order. Loci not included in the framework order were added to the order by reducing the stringency required for locus addition.

Linkage groups were combined to form a global map. Relative positions and orientations of linkage groups were determined using data from both BovR5 and BovR12 by fixing the established framework order in one linkage group and adding loci from a second linkage group using RHmaxlik with odds of 1,000:1. The locus order for the global map was fixed and inter-locus distances were calculated for both BovR5 and BovR12.

4.3 RESULTS

4.3.1 Chromosome microdissection library

Nine DNA sequences varying in length between 437bp and 735bp were obtained from a library of fragments from BTA7q14-q21. A very high level of primer artefact was observed in the sequences. Two of the sequences appeared to consist entirely of primer with no evident bovine DNA. The other seven sequences all contained a stretch of non-artefactual DNA followed by a stretch of primer artefact. The longest stretch of non-artefactual sequence obtained was 198bp. The sequences, with primer artefact removed, are shown in **Appendix 4.A**.

After removal of primer artefacts, sequences were checked for similarity to existing sequences in GenBank using the BLAST® tool (Altschul *et al.* 1997). One sequence was found to contain a 173bp stretch with 100% identity to the M13 bacteriophage database sequence (**Appendix 4.A**, Clone 23). A second sequence contained a 131bp stretch with extensive similarity (~80%) to the human Line-1 element (**Appendix 4.A**, clone 13). These sequences indicate contamination of the library with non-bovine genetic material. Further evidence of contamination of the library by extraneous DNA comes from a sequence fragment obtained in a separate laboratory. This 52bp sequence shows 100% identity to part of exon 2 of the bovine (also ovine and caprine) CYP19 gene (T. Goldammer, personal communication). CYP19 maps to bovine chromosome 10, and was cloned in the laboratory where the BTA7 library was constructed (Goldammer *et al.* 1994).

Of the five remaining sequences, one contained a 129 bp stretch showing very high similarity (~87%) to a bovine transposable element of the SINE (short interspersed element) family (**Appendix 4.A**, clone 19). A further four fragments out of six sequenced

(T. Goldammer, personal communication) also contained stretches of bovine transposable elements; three SINE family elements and one LINE (long interspersed element) family. These sequences are very common in the bovine genome, so it is unclear if they represent contamination by material from outside the targeted region of BTA7.

Four sequences show no significant similarity to any sequence in GenBank (**Appendix 4.A**, Clones 10, 12, 17, 18). Mapping would reveal whether these are sequences from Bta7q14-q21. However, the short length of the sequences (63bp to 152 bp) makes it unlikely that they will contain SNPs, hence no further work was done to characterise these sequences.

4.3.2 Novel SNP loci in genes

Fifteen genes were chosen to develop new SNP markers within the candidate genomic region on BTA7. In human, all 15 target genes had been radiation-hybrid mapped to the centromeric half of the p arm of chromosome HSA19 (**Table 4.1**) (Deloukas *et al.* 1998). Seven of these genes had been assigned to BTA7, of which six had been assigned positions on the chromosome by radiation hybrid mapping (**Table 4.2**).

Gene symbol	and description	GenBank Accession	Position (cR_{3000})*
DNMT1	DNA (cytosine-5-)-methyltransferase 1	NM_001379	67.48
ICAM5	Intercellular adhesion molecule 5, (telencephalin)	NM_003259	67.63
LDLR	Low density lipoprotein receptor	NM_000527	68.64
CALR	Calreticulin	NM_004343	71.27
MANB	Lysosomal alpha-mannosidase	AH006687	73.69
GCDH	Glutaryl-Coenzyme A dehydrogenase	NM_000159	73.9
LYL1	Lymphoblastic leukaemia derived sequence 1	M22638	74.42
CD97	Leukocyte antigen cd97	NM_001784	77.04
AKAP95	A-kinase anchoring protein	NM_005858	86.05
NOTCH3	Notch (Drosophila) homologue 3	NM_000435	86.05
JAK3	Janus kinase 3	NM_000215	98.11
RAB3A	RAS oncogene family member	NM_002866	103.08
CSPG3	Chondroitin sulphate proteoglycan 3 (Neurocan)	NM_004386	104.45
COMP	Cartilage oligomeric matrix protein	NM_000095	105.54
FKBP38	FK506-binding protein 8	NM_012181	124.01

Table 4.1: Human genes from HSA19 targeted for development of novel bovine SNPs
*Position on the GeneMap 99 radiation hybrid map (Deloukas *et al.* 1998).

Gene	Mapping	Reference
CALR	Synteny	Gao and Womack (1997)
COMP	RH	Gu *et al.*(1997), Gu *et al.*(1999)
DNMT1	RH	Kang'a *et al.* (2000)
LDLR	RH, Linkage	O'Brien *et al.* (1993), Gao and Womack (1997), Kang'a *et al.* (2000)
LYL1	RH	Gu *et al.*(1997), Gu *et al.*(1999)
MANB	RH	Gu *et al.*(1997), Gu *et al.*(1999)
RAB3A	RH	Gu *et al.*(1997), Gu *et al.*(1999)

Table 4.2: Human chromosome 19 genes from Table 4.1 mapped to chromosome BTA7
In column 2, 'RH' indicates that a locus has been mapped using radiation hybrid mapping. 'Linkage' indicates that a locus has been mapped using genetic linkage mapping. 'Synteny' indicates that a locus has been assigned to a chromosome using a somatic cell hybrid panel.

For four genes (MANB, RAB3A, LYL1 and COMP), published intronic bovine STS sequences (short, mapped DNA sequences) were available (Gu *et al.* 1997). The published primers were used to amplify these loci. For two of these four genes, and the remaining eleven genes, candidate introns were identified in human for amplification in cattle. Introns were selected for amplification on the basis of their length in human. Although the length

of a given intron varies between vertebrates (K. H. Wolfe, personal communication) it was assumed that the bovine intron would be comparable in length to its human orthologue. Care was taken to ensure that the human introns did not contain excessively long repetitive elements, although two introns did contain human SINE/Alu-type repeats, one them adjacent to a TA repeat. A further two introns contained GA-rich or C-rich low complexity regions (**Table 4.3**). However, in these four cases the repeat sequence accounted for less than half of the total human intron length.

Gene	GenBank accession	Intron start (bp)	Intron end (bp)	Intron size (bp)	Repetitive elements	Amplicon size in cattle (bp)
AKAP95	AC005785	21,271	21,715	445	-	450
CALR	AD000092	86,201	86,621	421	-	430
CD97	AC005327	13,644	14,073	430	-	N/A
COMP	AC003107	25,173	25,704	532	-	N/A
CSPG3	AC003110	27,922	28,464	543	C-rich 133bp	N/A
DNMT1	AC010077	14,561	15,005	445	$(TA)_n$ 52bp; SINE/Alu 128bp	~330
FKBP38	AC005387	24,662	25,135	474	-	~500
GCDH	AD000092	42,957	43,444	488	-	N/A
ICAM5 intron A	AF082802	4,889	5,339	451	-	N/A
ICAM5 intron B	AF082802	7,254	7,980	727	-	N/A
JAK3	U70065	7,504	7,917	414	GA-rich 113bp	~800
LDLR	AF217403	16,017	16,662	646	SINE/Alu 232bp	N/A
LYL1	AC005546	7,390	8,311	922	-	N/A
NOTCH3	AC004663	30,496	31,037	542	-	~1300†

Table 4.3: Human intron sequences from 13 HSA19 genes chosen for amplification in cattle
Accession numbers are those of the human genomic sequences used to identify candidate introns. Intron positions and sizes and repetitive elements relate to the human sequences. Amplicon size in cattle includes both flanking sequence and intron. For loci failing to amplify, size in cattle is indicated as N/A. †PCR product in human control sample was twice the expected size.

PCR products were obtained for six putative introns (**Table 4.3**) using bovine DNA samples and a human sample as a positive control. One locus (Notch3) gave a PCR product for the human control that was approximately twice the expected size. This suggests that the intended intron had not been amplified. It is possible that the product was part of a

paralogue of the Notch3 gene. For the remaining five loci, the human control gave a product of expected size. For these five loci, the bovine PCR product was considerably larger than would be obtained from flanking exon sequence alone, indicating that the bovine amplicons contained introns. Intron sizes for the loci in cattle vary between approximately 75% and 200% of the size in human. Primer sequences for the loci are given in **Table 4.5**.

Locus	Forward primers	Reverse primers
CD97	CTG CCC GGA GCT GCA CT GCT CCA CCC ACT GCC TCA ACA	GGA GCT GTC ACA CTG ATG CTG CGG ACA AAA CCC ATG CCA C
COMP	CCA CAG CAG CCG CAC CT GTC ATG TGG AAG CAG ATG GA	ACC ACA GCT CCA GCT TCT ATG T GAC GTA CTG GCA GGC GAA C
GCDH	GTG CCA GAG GAG AAT GTG C TGT CGA GGG CGT ACT GC	CAT CAT GGA CGG TGT GGA AAG CAG AAC TCC GAA GCT C
ICAM5 A	GCT CTC AGG CAC TTA CCG GCT TCT GTT CCC TCC AGC	CAA GGC GAG GCG GTC AA AGT AAT GCG TTC TGG GCA G
ICAM5 B	CCT CCA GCC CAG ATC AGC GGC CTC CTG CAC GTT GTA C	CGC CCA GGA GGA AAC TTC AC GGC CTC GCC TGA GCT CTC
LDLR	AGT GGC CCA ATG GCA TC TGA GAT GGA GTG AAG TTT GGA	CCC AAT GGC ATC ACC CTA G GTG AAG TTT GGA GTC AAC CCA
LYL1	GAA GCC GAT GTA CTT CAT GGC GTT GAA GCG GAG ACC AAG C	CGG AGC ACC TCG TTC TTG C CAA GCC ACT GTG AGC TGG AC
NEUR	ACC CAC AGG AAC TCT ACG ATG T TGT CCC ACC GAG GCC AG	ACT CTA CGA TGT GTA TTG CTT TGC CCG GGC CCA CGT AGA AG

Table 4.4: Primer sequences for eight putative bovine intron loci failing to amplify with PCR

Eight of the loci failed to give a clean PCR product in cattle (**Table 4.3**). For two of the loci (LDLR and COMP), no PCR product was seen, only primer dimer or a smear of background amplification artefact. For loci ICAM5 B and LYL1, there was weak amplification of a product similar in size to that expected if the intron had been lost. For the remaining four loci (CD97, CSPG3, GCDH, ICAM5) multiple PCR products of widely differing sizes were obtained. No further work was carried out with these eight loci. Failure to amplify the intended target demonstrates the difficulty of PCR using primers designed in a different species. Primers for the eight loci are given in **Table 4.4**.

Locus	Forward primer	Reverse primer
AKAP95	CTG TAG CTG GGG CGG TAG	GGA GGG CAT ACA GGA CCG
CALR	GGC TAT GTG AAC GTG TTT CCA G	AGT TGA AGA TGA CAT GAA CCT TC
DNMT1	GGC CCA AGA TTT TTG CCA T	AAT ATC GAA CTC TTC TTT TCT GG
FKBP38	GAG CTC TGC GTG GAT CGT	GCC CTG AAG CTG GAA CCT
JAK3	TGG AAG CAG CGA GCT TGA	CCA GGT GTA CAA ATT CCT GC
Notch3	CAG CGT GCC CTC AAA GC	CTA CGA GTG CCG CTG TGC
RAB3A*	ACT CTC CCC CAA CCA TGA A	CAG GGC AGG TAG TAC CAG TGA T
MANB*	AGC CCA GCT TCT ATC TCT AGC AC	AGA AGC CGT CAA AAC CCA TC

Table 4.5: PCR primers for amplification of BTA7 intron loci

*Previously published primer sequences (Gu *et al.* 1997).

In addition to the novel intron sequences, the previously published STS loci MANB and RAB3A were successfully amplified.

4.3.3 Sequencing

Sequences were obtained for the five novel introns successfully amplified (**Table 4.3**) and the two STS loci MANB and RAB3A. For each locus, three samples were sequenced, one from each of three different breeds representing the principal bovine lineages: Ongole (pure-bred *B. indicus*); Charolais or Jersey (pure-bred European *B. taurus*); Guinean N'Dama (pure-bred African *B. taurus*). Guinean N'Dama have been shown to be the purest West African taurine population, with less than one percent genetic introgression from zebu (MacHugh *et al.* 1997). Ongole are the purest Indian zebu breed genetically characterised (P. Kumar, in preparation).

For loci MANB, RAB3A, JAK3 and CALR products were sequenced in both forward and reverse directions. For the remaining three loci, only the forward sequence was obtained. For each locus, the three cattle sequences were aligned and a consensus sequence derived. For the five novel intron sequences, the consensus sequence was aligned using ClustalX (Jeanmougin *et al.* 1998) with the orthologous human sequence from GenBank, for which

the intron position was known. Exon fragments at the start and end of bovine sequences were identified and translated. Translated bovine exons were aligned with the human and mouse protein sequence (**Figure 4.2**).

Locus	Species	Position in protein	Amino acid sequence
AKAP95	Human	123	ES -Intron- SSFRFQPFESYDS
	Cattle		.. S.....S.....
	Mouse		D. S......Y....A
CALR	Human	130	IMF -Intron- GPDICGPGT
	Cattle		... .
	Mouse		... .
DNMT1	Human	438	SAKPIYDDDPSLE -Intron-
	Cattle		......E....P.
	Mouse		C..A.H.EN..M.
FKBP38	Human	281	SNK -Intron-
	Cattle		...
	Mouse		...
JAK3	Human	577	SQVSYRHLVLLHGVCMAGD -Intron-
	Cattle		.....Q.....
	Mouse		.....P.....

Figure 4.2: Protein sequence alignments of exonic regions of five novel BTA7 amplicons
Numbers indicate the position of the first amino acid residue in the alignment in the human protein sequence. Intron positions in the human sequence are shown. Identity to the human amino acid residue is represented by a dot.

For each of the five novel BTA7 sequences, the protein sequence alignments indicate that the bovine sequence is the orthologue of the targeted human gene. For loci CALR and AKAP95, the cattle sequence spans the entire target intron, with regions of flanking exon both before and after the intron. For loci DNMT1, FKBP38 and JAK3, the cattle sequence starts in the first exon but stops before the end of the subsequent intron, as the sequencing

primer used was too close to the start of the second exon to allow sequencing of the entire intron.

4.3.4 Sequence polymorphism

For the seven loci sequenced in N'Dama, Ongole and Charolais or Jersey, only JAK3 was non-polymorphic. For the other six loci, a total of 18 SNPs were found in the three animals (**Table 4.6**). All 18 SNP sites have two alleles. 14 of the SNPs are transitions, and 4 are transversions. Two SNPs are in exon sequence (one in DNMT1 and one in MANB), but both are at third codon positions and neither changes the amino acid sequence. The total length of DNA sequenced in the three animals was 4699bp, of which 388bp is exon sequence and 4311bp is intron sequence. The overall frequency of nucleotide variation is 1 SNP per 261bp, and overall nucleotide diversity is $1.7 \pm 1.4 \times 10^{-3}$. For intron sequence there is 1 SNP per 269bp, and for exon sequence, 1 SNP per 194bp.

Of the 18 SNP loci found, 11 have an allele seen only in the Ongole (*Bos indicus*). It is possible that these alleles are unique to zebu animals. Previous research using 20 microsatellites has shown that half of the loci display alleles that are present at high frequency (~67%) in zebu and absent or very rare (<1%) in pure taurine populations (MacHugh *et al.* 1997). The interpretation is that the alleles are of zebu origin, and that their presence at low frequency in taurine populations is due to admixture. Similarly, a study of two genes sequenced in a panel of taurine and indicine cattle showed a much higher number of polymorphic bases in the indicine than the taurine animals, suggesting that many alleles were unique to *B. indicus* (Konfortov *et al.* 1999).

Three SNPs have an allele exclusive to one or other of the *Bos taurus* breeds of which two are unique to N'Dama, and one is unique to Charolais (**Table 4.6**). Over all loci, pairwise

nucleotide diversity (average number of nucleotide differences per site) is greater between the Ongole and the two taurine samples ($2.1x10^{-3}$ and $2.8x10^{-3}$) than between the two taurine samples ($2.1x10^{-4}$).

At nine of the 18 SNP loci, the Ongole is homozygous for one allele and the two pure taurine animals homozygous for another, suggesting that one allele might be fixed in zebu animals and the other in taurine (**Table 4.6**). The West African N'Dama is heterozygous at five of the SNP sites. This compares with two heterozygous sites in Ongole, and one heterozygous site in Charolais.

Gene	FKBP38			RAB3A		AKAP95	CALR					MANB					DNMT1				
Base / Sequence	176	307	431	193	576	270	108	141	160	190	317	41	206	446	1116	1123	28	105	125	144	248
European*	C	C	T	C	C/T	T	A	A	G	T	T	T	G	C	C	T	C	T	T	A	C
Guinean N'Dama	C	C	T	C/T	C	T	A/G	A/G	C/G	T	T	G/T	G	C	C	C	C	T	T	A	C
Ongole	C	C	C	C	T	C	G	G	G	C	C	G	A	C/T	C	C	T	C	C	A/C	G
ILRI N'Dama 1	C	C	T	C/T	C	T	A/G	A/G	G	T	T	T	G	C	C/T		C	T	T	A	
ILRI N'Dama 2	C	C	T	C/T	C	T	A/G	A/G	G	T	T	T	G	C	T		C/T				
ILRI N'Dama 5	C	C	T	T	C	T	A	A	G	T	T	T	G	C	T		C				
ILRI N'Dama 6	C/T	C	C/T	C/T	C/T	T	A/G	A/G	G	T	T	T	G	C	C/T		C				
ILRI Boran 3	C	C	T	C	C	C	A	A	G	T	T	T	G	C	C/T		T				
ILRI Boran 4	C	C	C	C		C/T	A	A	G	T	T	T	G	C	C/T		T				
ILRI Boran 7	C	C/T	C	C	C/T	C	A/G	A/G	G	C/T	C/T	T	A/G	C	C/T		C/T				
ILRI Boran 8	C	C	T			C/T	G	G	G	C	C										

Table 4.6: Polymorphic bases in six BTA7 sequences typed in individuals from cattle from Europe, Africa and India and a hybrid African *Bos taurus* - *Bos indicus* pedigree

*A Jersey DNA sample was used to obtain the sequence for locus CALR. All other European sequences were obtained using a Charolais DNA sample.

Numbers show positions of polymorphic bases in the sequence (**Appendix 4.B**). Homozygotes are denoted by a single base character and heterozygotes by two characters. SNPs shaded in dark grey are those where the Ongole is homozygous for one allele and the two pure taurine animals homozygous for the other. Blank cells shaded in light grey indicate bases not typed in a particular animal.

4.3.5 SNP typing in ILRI trypanotolerance-mapping pedigree

The founder animals of the ILRI trypanotolerance gene mapping pedigree were sequenced to determine whether the SNPs identified were polymorphic in the pedigree. The pedigree was constructed using eight founder animals: four longhorn *Bos taurus* N'Dama from the Gambia in West Africa and four *Bos indicus* Boran animals from Kenya. It was not possible to type all eight founder animals for all of the polymorphic sites. One sample (Boran 8) proved particularly difficult to sequence, probably due to DNA degradation. Data was obtained for 14 of the SNPs from six different intron loci (**Table 4.6**). Eleven SNPs were polymorphic in the ILRI founder animals and three were monomorphic. Three additional SNPs were found in individual pedigree founder animals; two at locus FKBP38 and one at locus MANB (**Table 4.6**). All three are biallelic transitions occurring in intron sequence.

Whereas sequencing of pure-bred animals suggests that many of the SNPs have 'diagnostic' zebu or taurine alleles (**Table 4.6**), it is not possible to distinguish clearly on the basis of SNP genotype between the two breeds used to construct the ILRI pedigree. There is no SNP where all Boran grandparents are homozygous for one allele, whilst all N'Dama grandparents are homozygous for the other.

Considering only those SNPs that were fixed for one allele in the Ongole and a different allele in the two Bos taurus animals, we constructed putative haplotypes for the ILRI

pedigree founder animals to investigate admixture at a sub-chromosomal level. Haplotypes consist of six SNPs spanning five genes (**Table 4.7**).

Gene: / Sequence	FKBP38	AKAP95	CALR		MANB	DNMT1	No. of obligate recombinations
Guinean N'Dama	T	T	T	T	G	C	0
Ongole	C	C	C	C	A	T	0
ILRI N'Dama 1	T	T	T	T	G	C	0
	T	T	T	T	G	C	0
ILRI N'Dama 2	T	T	T	T	G	C	0
	T	T	T	T	G	T	1
ILRI N'Dama 5	T	T	T	T	G	C	0
	T	T	T	T	G	C	0
ILRI N'Dama 6	C	T	T	T	G	C	1
	T	T	T	T	G	C	0
ILRI Boran 3	T	C	T	T	G	T	3
	T	C	T	T	G	T	3
ILRI Boran 4	C	C	T	T	G	T	2
	C	T	T	T	G	T	2
ILRI Boran 7*	C	C	C	C	A	T	0
	C	C	T	T	G	C	1
ILRI Boran 8	T	C	C	C			1
	T	T	C	C			1

Table 4.7: BTA7 SNP loci haplotypes of ILRI trypanotolerance trait-mapping pedigree founder animals

For deduction of gene order see details of map construction below.

SNPs shown are those in **Table 4.6** (shaded dark grey) for which Ongole and pure taurine animals are homozygous for different alleles, and which were typed in the ILRI founder animals.

Putative zebu alleles are shaded in dark grey, taurine alleles in light grey. Two SNPs were not typed for Boran 8 and are indicated as blank, white cells.

*Haplotypes for Boran 7 are ambiguous. The two haplotypes shown are those requiring the lowest number of recombinations.

For the six SNP loci in the haplotypes, we assume that the base in the Ongole is fixed in zebu cattle, and that the base seen in the Guinean N'Dama is fixed in African taurine cattle.

Although only one purebred animal from each sub-type of cattle was sequenced, this assumption does not appear unreasonable in the light of previous research into zebu and taurine-specific alleles, as discussed above (Konfortov *et al.* 1999; MacHugh *et al.* 1997). Considering the haplotypes of the ILRI pedigree founder animals, it is possible to determine whether each base in each haplotype is taurine or zebu in origin, and to count the number of required recombination events. Under the stated assumptions, the four Gambian N'Dama would appear to show a low level of zebu / taurine genetic admixture, with only two 'zebu' alleles compared with 46 'taurine' alleles. Two of the eight haplotypes appear to be recombinant, with one recombination required in each case (N'Dama 2 and N'Dama 6). In contrast, the four Boran animals would seem to show substantial admixture. Of 44 bases typed, 22 show the taurine allele and 22 the zebu allele, indicating equal zebu and taurine genetic contributions to the breed. There appears to be a high level of historical recombination. Of eight haplotypes, only one is inferred to be non-recombinant, with the other seven requiring from one to three recombination events. Recombination events are inferred in three of the four inter-genic intervals.

4.3.6 Somatic cell hybrid mapping of novel loci

The unmapped genes AKAP95, FKBP38 and JAK3 were unambiguously assigned to BTA7 by somatic cell hybrid (SCH) mapping, with correlation coefficients of 0.97, 0.97 and 0.93 respectively. The previous assignment of gene CALR to BTA7 was confirmed with a correlation coefficient of 0.97 (Johnson 2000). The product obtained using primers designed for locus NOTCH3 showed highest concordance for chromosome 21 (correlation coefficient 0.92) (Johnson 2000). This further suggests that the product amplified was not the intended NOTCH3 intron.

4.3.7 Radiation hybrid mapping

To refine the radiation hybrid map of the candidate trypanotolerance region, new loci were scored for presence or absence in two bovine / Chinese hamster whole genome radiation hybrid panels of different resolution. The first panel (BovR5) consisted of 90 hybrids given a total dose of 5000 rad (Womack *et al.* 1997). The second panel (BovR12) consisted of 180 hybrids treated with a higher intensity dose of 12000 rad, yielding smaller chromosome fragments (Rexroad *et al.* 2000). Typical radiation hybrid PCR results are shown in **Figure 4.3**.

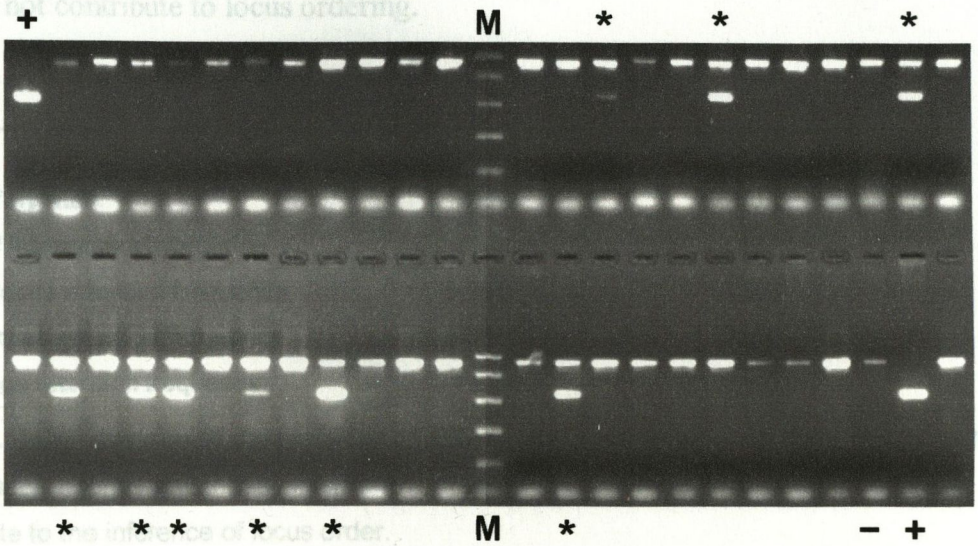

Figure 4.3: Radiation hybrid Calreticulin (CALR) gene PCR for 45 BovR12 hybrids and controls run in a 2% agarose gel and stained with EtBr

Lanes marked 'M' contain a molecular size standard with bands of 50, 150, 300, 500, 750 and 1,000bp; Positive controls (Aberdeen Angus) are marked '+', and show only the bovine CALR PCR product; a negative control (Chinese hamster) is marked '-' and shows only the hamster CALR PCR product; Other lanes contain CALR PCR products for 45 different radiation hybrid lines. Lanes marked with an asterisk are hybrids retaining the bovine CALR gene.

Ten loci were added to the radiation hybrid map. These included the four novel intron loci previously unmapped in cattle, and six microsatellites previously mapped on genetic linkage maps (Gu *et al.* 2000). Four of the microsatellites had previously been typed in

panel BovR5 (Kang'a *et al.* 2000), and only required typing in BovR12. Data for the ten loci typed was added to an existing dataset of loci typed at ILRI (Kang'a *et al.* 2000). This gave a total dataset of 31 loci typed in both panels of which 19 are genes and 12 are microsatellites. An additional two loci were typed only in BovR12.

Retention frequencies for the two panels are very similar. However, information content is much greater for panel BovR12 (**Table 4.8**). BovR5 contains half as many hybrids as panel BovR12, and a higher percentage of the BovR5 hybrids are non-informative (58% compared with 43% for panel BovR12) in that they retain all or none of the loci typed and thus do not contribute to locus ordering.

	BovR12	BovR5
No. of Hybrids	180	90
No. of informative hybrids*	103	41
Max. locus retention frequency	0.194	0.189
Min. locus retention frequency	0.085	0.070
Average retention frequency	0.131 ± 0.006	0.133 ± 0.006

Table 4.8: Comparison of datasets for 31 BTA7 loci typed in two radiation hybrid panels
*Informative hybrids are those which show at least one inter-locus chromosome break, and thus contribute to the inference of locus order.

4.3.8 BTA7 loci linkage groups

Two-point analysis was used to test linkage for all pairwise combinations of loci and thus define groups of linked loci for subsequent ordering. Pairwise linkage is high for the 31 loci typed in the BovR5 panel. Of 31 loci typed, 27 are linked at Lod 7.0, with the remaining four loci unlinked. A lower level of linkage is seen in the BovR12 panel. At Lod 3.0, all 33 loci typed are linked. At Lod 4.0 or 5.0, two groups are defined: one group of 30 loci and a second of 3 loci. At Lod 6.0, the group of 30 loci breaks into three groups of 6, 14 and 10 loci, while the group of 3 loci remains intact. These four linkage groups were

designated Groups 1 to 4 (**Table 4.9**), and were used for subsequent mapping. Further two-point analysis with increasing Lod score shows group 3 to be the most tightly linked. Nine of the ten loci remain linked at Lod 14.0. Linkage is also tight in group 1, where four of the six loci are linked at Lod 14.0.

Linkage group	Group 1	Group 2	Group 3	Group 4
Loci	JAK3	AKAP95	IL5	BM7247*
	RM012	BL5	IL13	UWCA20
	FKBP38	BP41	IRF1	S14
	COMP	BL1067	IL4	
	ILSTS001	CALR	IL3	
	BMS713	MANB	CSF2	
		DNMT1	K21	
		ICAM3	S10	
		LDLR	S46*	
		EPOR	BM6105	
		NDUFA7		
		ARP(NG27)		
		RM006		
		MYO1F		

Table 4.9: Linkage groups for 33 BTA7 loci typed in a BovR12 rad radiation hybrid panel
Linkage groups were obtained using panel BovR12 data at a Lod score of 6.
Underlined loci are type I loci (genes). Loci typed only in BovR12 are marked with an asterisk (*).

4.3.9 Choice of model for radiation hybrid mapping

For ordering loci, it is necessary to select the appropriate model for the relationship between chromosomal location and locus retention probability. Gu *et al.* (1999) found that locus retention frequencies for 32 loci from the whole of BTA7 typed in panel BovR5 conform to the equal retention model, where retention is independent of chromosomal position (Boehnke *et al.* 1991). This model has also been found appropriate in mapping studies using BovR5 of chromosomes BTA1 (Rexroad *et al.* 1999), BTA5 (Ozawa *et al.* 2000) and BTA23 (Band *et al.* 1998), and in a study of BTA1 using the BovR12 panel

(Rexroad *et al.* 2000). An alternative model, where retention probability is increased for fragments containing the centromere or telomere (Lawrence *et al.* 1991) was used in constructing a BovR5 panel map of BTA13 (Schlapfer *et al.* 1997).

For the loci typed here, average fragment retention probabilities are very similar for both panels (**Table 4.8**). However, there is no significant correlation between retention frequencies in the two panels (Pearson 'r' coefficient 0.22, 29 d.f.) (**Figure 4.4**), indicating that there is no consistent relationship between the position of loci on the chromosome and their retention probability. This suggests that the equal retention probability model is appropriate.

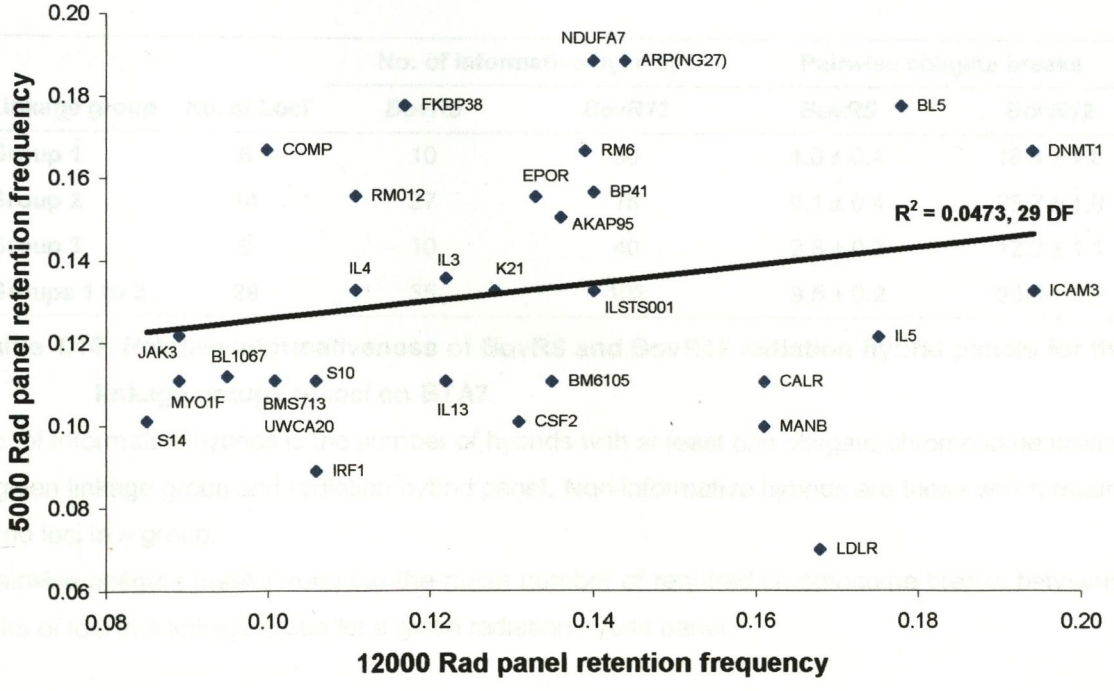

Figure 4.4: Correlation of locus retention frequencies for two radiation hybrid panels
Retention frequencies are calculated as the proportion of lines retaining a given genetic marker in each panel.

4.3.10 Locus order determination

Maximum likelihood multipoint analysis was applied (where possible) to each of the four linkage groups (**Table 4.9**) in each RH panel to identify 'framework loci' (Lange *et al.*

1995). Framework loci are those incorporated into an order such that the ratio of likelihoods for the best and next best orders is at least 1,000:1. Panel BovR12 performs better than BovR5; for linkage groups 1 to 3, more loci were ordered at odds of 1,000:1 using BovR12 than BovR5 (**Table 4.11**). For groups 1 and 3, it was not possible to order any loci at odds of 1,000:1 using panel BovR5. The inability of panel BovR5 to resolve locus order within linkage groups reflects the low information content of the panel compared with BovR12. For linkage groups 1 to 3, a far higher number of BovR12 hybrids are informative than BovR5 hybrids (**Table 4.10**). Counting obligate chromosome breaks for all possible pairs of loci in each linkage group similarly reveals a significantly higher level of chromosome breakage in the BovR12 panel.

		No. of informative hybrids		Pairwise obligate breaks	
Linkage group	**No. of Loci**	***BovR5***	***BovR12***	***BovR5***	***BovR12***
Group 1	6	10	36	4.0 ± 0.4	16.3 ± 1.8
Group 2	14	27	78	9.1 ± 0.4	25.7 ± 1.0
Group 3	9	10	40	2.8 ± 0.3	12.2 ± 1.1
Groups 1 to 3	29	35	102	9.5 ± 0.2	29.3 ± 0.5

Table 4.10: Relative informativeness of BovR5 and BovR12 radiation hybrid panels for three linkage groups of loci on BTA7

No. of informative hybrids is the number of hybrids with at least one obligate chromosome break for a given linkage group and radiation hybrid panel. Non-informative hybrids are those which retain all or no loci in a group.

'Pairwise obligate breaks' refers to the mean number of required chromosome breaks between all pairs of loci in a linkage group for a given radiation hybrid panel.

Multipoint analysis was also performed using both panels simultaneously. In this approach, the likelihood for each locus order is calculated independently for each panel, and likelihoods from the different panels are multiplied to give an overall likelihood (Lunetta *et al.* 1996).

Where multiple panels are used, it is necessary to decide whether to treat distances in the different panels proportional or non-proportional. Inter-locus distances obtained for loci on bovine chromosome BTA1 using the BovR12 and BovR5 panels are non-proportional (Rexroad *et al.* 2000), hence it was assumed that inter-locus distances in the BovR5 and BovR12 panels for BTA7 would also be non-proportional.

Simultaneous multi-point analysis of both RH panels gives framework maps as good or better than those obtained by independent analysis of the two panels (**Table 4.11**). For linkage group 2, the improvement in the map is dramatic, with all 14 loci ordered at odds of 10^5:1. For group 3, six loci are ordered at odds of 10^6:1 compared with just three framework loci for BovR12. Due to the increased statistical confidence of the framework maps constructed using both panels simultaneously, both panels were used to construct the final map.

	Number of framework loci (ordered at odds of >=1,000:1)		
Linkage group	Panel BovR5	Panel BovR12	Both Panels
Group 1 (6 loci)	0	4	4
Group 2 (14 loci)	4	8	14
Group 3 (9/10 loci)*	0	3	6
Group 4† (3 loci)	-	3	-

Table 4.11: Efficacy of ordering loci using BovR5 and BovR12 radiation hybrid panels independently and in combination

* 9 group 3 loci were typed in the BovR5 panel and 10 in the BovR12 panel.

† Only two loci from group 4 were typed in panel BovR5, preventing ordering of loci using the panel.

4.3.11 Integration of linkage group framework maps

Groups 1 and 2 were integrated by multipoint analysis using data for the 20 loci in both groups. A framework order of 18 loci ordered at odds of 2000:1 was obtained. In the

framework, all 14 group 2 loci and the four group 1 framework loci were present in the same order as previously found.

Integrating group 3 into the map was more problematic. Fixing the order of loci in groups 1 and 2 and adding group 3 loci to the growing framework or vice versa shows that groups 1 and 3 lie on opposite sides of group 2 (odds > 10^5:1). The orientation of group 3 with respect to group 2 is less clear. By fixing the order of loci in either group 3 or groups 1 and 2, and adding the remaining loci, one orientation is favoured at odds of approximately 400:1. This orientation is the same as that seen in the BTA7 framework (odds of 1,000:1) linkage map (Gu *et al.* 2000), and is therefore the orientation used in the final map (**Figure 4.5**).

Finally, group 4 was added to the map. Two-point analysis had revealed group 4 to be the least closely linked to the other groups. Multi-point analysis does not give a good indication of the position or orientation of group 4 relative to the other groups, though it does show that group 4 lies at an end of the map. Group 4 was therefore placed on the final RH map by referring to the position of group 4 loci on the comprehensive linkage map of Gu *et al.* (2000). The linkage map contains two of the three group 4 loci; BM7247 and UWCA20. Of these, only UWCA20 is a framework locus in the linkage map, and it is not clear whether BM7247 lies on the proximal or distal side of UWCA20. For this reason, even though all three group 4 loci are ordered with high confidence (BM7247 - UWCA20 - S14, odds 50,000:1), only one is considered to be a framework locus in the final map.

4.3.12 Comprehensive map

The comprehensive map contains 33 loci, of which 19 are genes. Linkage group 1 is at the centromeric end of the map, followed in succession by groups 2, 3 and 4 (**Figure 4.5**).

Using the established locus order, inter-locus distances were calculated for both panels according to the formula:

$$d = -\ln(1-\theta)$$

Equation 4.1

where d is additive distance between a pair of loci and θ is the probability of a radiation-induced break between the loci (Boehnke *et al.* 1991). Using the BovR12 data (33 loci), the established locus order requires 404 obligate chromosome breaks. The length of the map is 1306.7 cR_{12000}. The BovR5 data (31 loci) gives a total length of 698.6 cR_{5000}, with 121 obligate chromosome breaks.

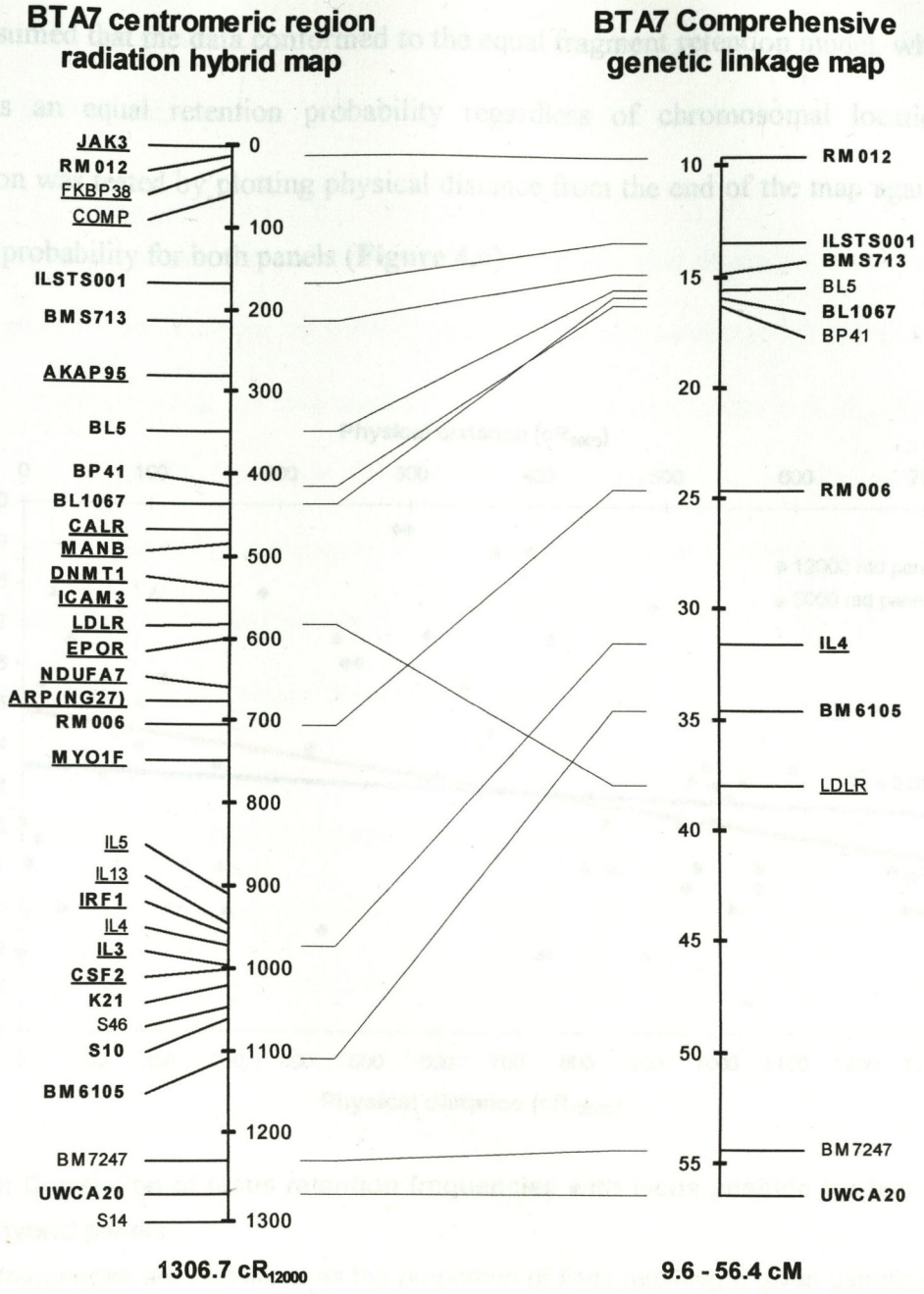

Figure 4.5: Comparison of radiation hybrid and genetic linkage maps of the centromeric region of bovine chromosome 7

The two-panel radiation hybrid map is on the left and the comprehensive genetic linkage map of Gu *et al.* (2000) is on the right. Locus symbols in bold type are framework loci ordered at odds of 1,000:1 or greater. Underlined symbols are Type I loci (genes)

4.3.13 Validation of model assumed in creating radiation hybrid maps

It was assumed that the data conformed to the equal fragment retention model, where each locus has an equal retention probability regardless of chromosomal location. This assumption was tested by plotting physical distance from the end of the map against locus retention probability for both panels (**Figure 4.6**).

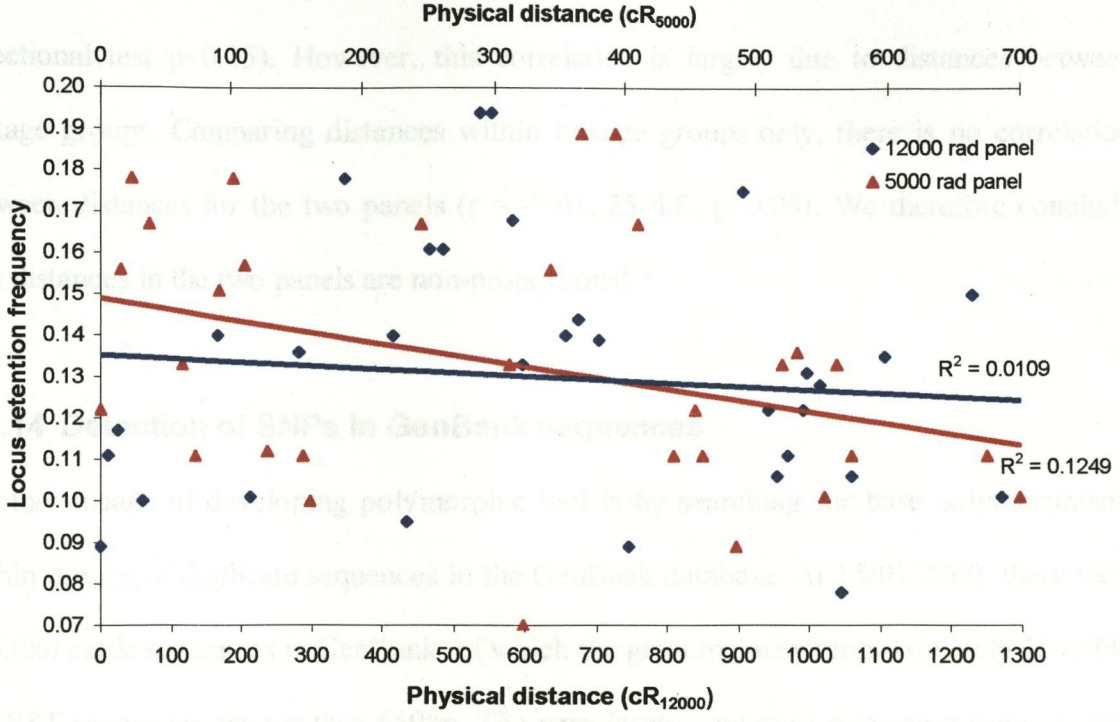

Figure 4.6: Correlation of locus retention frequencies with locus position for two radiation hybrid panels

Retention frequencies are calculated as the proportion of lines retaining a given genetic marker in each panel. Distances are centi-rad distances from the end of the radiation hybrid map derived using data from both panels.

For panel BovR5, physical distance from the centromere (in cR_{5000}) is negatively correlated with locus retention frequency (**Figure 4.6**). However, the correlation is significant only at the level of $p<0.05$ (Pearson 'r' coefficient -0.35, 29 d.f., one-tailed test), and there is no correlation between distance from the centromere and retention frequency within any of the linkage groups. For BovR12, there is no significant correlation

between position and retention frequency, either globally across all linkage groups (**Figure 4.6**), or within any of the groups. We therefore conclude that the equal-retention model adopted is appropriate.

A second assumption made in order to create the map was that distances in the two panels were not proportional. Comparing (non-cumulative) inter-locus distances for all loci in the two panels shows a weak positive correlation (Pearson r coefficient = 0.34, 28 d.f., one-directional test $p<0.05$). However, this correlation is largely due to distances between linkage groups. Comparing distances within linkage groups only, there is no correlation between distances for the two panels ($r = -0.01$, 25 d.f., $p>0.05$). We therefore conclude that distances in the two panels are non-proportional.

4.3.14 Detection of SNPs in GenBank sequences

Another means of developing polymorphic loci is by searching for base polymorphisms within groups of duplicate sequences in the GenBank database. At 15/01/2000, there were 165,080 cattle sequences in GenBank, of which the great majority (approximately 161,000) are EST sequences shorter than 650bp. The very large number of sequences suggests that many genes will have been sequenced more than once, possibly using different bovine genetic source material. Groups of duplicate sequences can therefore be identified and surveyed for the presence of SNPs. A software application called 'SNP Hunter', written in the Perl programming language, was created to facilitate this process. Details on how to use SNP Hunter are given **Section 6.2,** and ā full listing of the program is given in **Appendix 6.A**.

All sequences in the database at 15/01/2001 were downloaded. Sequences over 1000bp in length were removed, leaving 163,215 sequences. Each of these was tested for similarity to

all other sequences using the BLAST® tool (Altschul *et al.* 1997). A very high level of stringency was used (expectation value 'E' set to $1.00x10^{-250}$) in order to find genuine duplicate sequences. 6,962 separate groups of duplicate sequences were obtained. A minimum of two copies of each allele were required to classify a base as polymorphic, so as to eliminate artefactual 'polymorphism' due to sequencing errors. Hence groups of fewer than four sequences were discarded, leaving 1,486 groups. For groups of more than 40 sequences, sequences were removed at random until only 40 remained. For each group, all sequences were aligned using ClustalW (Thompson *et al.* 1994). The first 40bp of each sequence in the alignment was ignored to reduce the risk of erroneously classifying sequencing artefacts as SNPs. The criterion for SNP classification was that there should be at least two alleles present in at least two copies. For groups of more than 10 sequences, this criterion was modified such that a frequency of at least 20% was required for the second commonest allele. 292 groups of duplicate sequences showed at least one SNP. Consensus sequences for these groups were obtained and tested for similarity to sequences in the NCBI human genome contig map using the BLAST® tool. Human orthologues of the bovine sequences could thus be identified. By reference to the genomic location of the human orthologue, it is possible through comparative mapping to predict the genomic location of the bovine sequence. For 239 of the bovine SNP sequences, apparent human orthologues were identified at an E value of $1.00x10^{-10}$. For 155 SNP sequences, the human orthologues were restricted to only one human chromosome.

Eleven loci show single orthologues on human chromosome 19. The BLAST® tool was used to search for similarity between these sequences and the HSA19 'Golden Path' contigs (7th Oct 2000 dataset, obtainable from http://genome.cse.ucsc.edu). Six loci have orthologues within the first 22Mb of the chromosome, in the region of conserved synteny with BTA7. Details of the six loci are shown in **Table 4.12**, and the sequences are included

in **Appendix 4.C**. Each sequence contains a single SNP site. Five of the SNPs are transitions and one is a transversion. For five of the loci, the corresponding human gene is characterised.

Sequence	Position (Mb)	E-value	Base	Polymorphism				Human gene
				G	A	T	C	
SNP_Seq1	0.7	3.0E-23	134			2	2	AB013891Inward rectifier potassium channel Kir7.1
SNP_Seq2	4.7	6.0E-31	279	6	5			None
SNP_Seq3	11.3	1.0E-69	174	6	6			XM_009030 Eukaryotic translation initiation factor 3, subunit 4 (EIF3S4)
SNP_Seq4	12.0	4.0E-72	177			8	4	XM_009080 Putative T1/ST2 receptor binding protein (IL1RL1LG)
SNP_Seq5	21.2	1.0E-34	256	2			2	AB011133 KIAA0561 protein
SNP_Seq6	22.0	7.0E-49	316	7	5			XM_009333 Coatomer protein complex, subunit epsilon (COPE)

Table 4.12: Six bovine SNP loci with human orthologues mapping to HSA19 centromeric region identified from GenBank sequences

'Position' indicates the approximate position of the human orthologue on HSA19 as determined from the 'Golden Path' contig map of HSA19. 'E-value' indicates the significance of the human-bovine BLAST® result. 'Base' is the number of the SNP in the bovine consensus sequence (shown in **Appendix 4.C**). 'Polymorphism' shows the alleles present in the bovine sequences at the SNP site.

4.4 DISCUSSION

This study shows that is possible to develop novel single nucleotide polymorphism (SNP) markers within targeted regions of the bovine genome, and to fine-map markers with high confidence using the bovine radiation hybrid panels available.

4.4.1 Methods for isolating novel genetic markers in cattle

The use of comparative mapping to locate new markers within a specific chromosomal region seems a more profitable approach than the use of chromosome microdissection libraries for a number of reasons. The BTA7q14-22 library used in this study appears to contain many contaminant sequences from outside the target bovine chromosome region. Of thirteen DNA sequences obtained from the library, two were clearly not from the target bovine chromosome region. A third sequence that was highly similar to a human repeat element may also represent a contaminant. This suggests contamination from at least three different sources. The contamination of the library is most likely due to limitations in the techniques used in constructing the library. Conventionally, PCR uses thousands of copies of the template DNA with highly specific primers to amplify only the intended DNA sequence. Contaminant DNA present at low copy number has little effect on PCR for two reasons. First, the very large number of copies of the correct template swamps any contaminant. Second, the PCR primers bind ineffectually in the event of mismatch between primer and target. In contrast, in constructing the BTA7 library, fewer than ten copies of the template were used, and non-specific primers were used to amplify sequences throughout the target region (Goldammer *et al.* 1996). In addition to the obvious contaminants, a number of unidentified sequences were obtained. These may genuinely have been bovine sequences from the candidate region. However, all were short, and five consisted largely of common repeat elements, so the sequences would probably not be

good templates for PCR. Three anonymous SNPs have been developed using the library and mapped to the candidate region on BTA7 (Kang'a *et al.* 2000) (**Figure 4.5**, loci K21, S46, S14). However, we believe that a comparative genomic approach offers better potential for isolating novel SNP sequences from specific chromosomal regions. Other workers have reported success in finding microsatellites using a bovine BTA6q21-31 library constructed using the same method (Weikard *et al.* 1997). Weikard *et al.* identified eleven sequences carrying microsatellites, of which six mapped to the intended sub-chromosomal target region. Comparative genomics cannot be used to isolate novel bovine microsatellites from a specific region, as these loci are not conserved between human and cattle. The BTA7 library might therefore have greatest potential for isolating novel microsatellites from the centromeric region of BTA7.

A further disadvantage of using a chromosome microdissection library to isolate novel markers is that the markers could derive from anywhere within the chromosome fragment, in this case one third of the chromosome. In contrast, use of comparative mapping information allows more markers to be isolated from much more narrowly defined chromosomal regions. The increasingly detailed bovine radiation hybrid map (Band *et al.* 2000) shows the relationships between the bovine and human chromosomes. It is therefore possible to identify which regions of the human genome correspond to a given bovine chromosome region. Radiation hybrid map locations are known for some 30,000 human genes (Deloukas *et al.* 1998), and advances in the human genome project mean that approximate physical positions in Mb are known for many genes (see e.g. http://www.ensembl.org/). There is thus a wealth of precisely mapped human genes available for developing novel bovine genetic markers. Mapping human genes in cattle will further elucidate the relationship of the bovine to the human genome map.

4.4.2 Use of bovine EST sequences in mapping and SNP identification

The creation of bovine expressed sequence tag (EST) libraries provides another means of exploiting comparative mapping information to obtain new markers for inclusion in the bovine map. ESTs are unique, short DNA sequences derived from cDNA libraries, and therefore from sequences which have been transcribed in some tissue or at some stage of development. As EST sequences are relatively easy to generate, these sequences now account for almost 98% of all bovine DNA sequences. Bovine ESTs can be identified by homology to human genes or ESTs, and their approximate position in the bovine genome can be predicted by reference to the human map (Band *et al.* 2000; Ma *et al.* 1998). Band *et al.* found that 95% of bovine ESTs with a recognisable human orthologue map to the chromosomes predicted by comparative mapping. Researchers can thus choose ESTs from genes likely to be present in a specific genomic region for constructing highly detailed maps.

In many cases, duplicate sequences have been isolated from different bovine source material, and it is possible to search for polymorphism within these duplicates. This *in silico* approach may offer a shortcut to identifying SNPs. Here, a total of 292 apparent bovine SNPs were identified using the bovine sequences available at 15/01/2001.

Caution must be used in identifying SNPs for a number of reasons. First, it is likely that there are sequencing errors in the database sequences. For this reason, bases were only considered to be polymorphic if at least two alleles were found in at least two copies. Also, regions close to the start and end points of sequences were discarded. A second problem is the existence of paralogous genes. These are members of the same gene family which share a common ancestor due to a historical duplication event. Subsequent chromosomal rearrangement may have occurred, causing paralogues to have different genomic locations.

If duplication and rearrangement occurred prior to the divergence of the human and bovine lineages, then we might expect paralogues to be present in both human and cattle. A member of a bovine gene family would then show high similarity to human sequences mapping to different locations. This is the case for at least 137 of the 292 bovine SNP sequences, which have apparent orthologues on more than one human chromosome.

After discarding apparent sequence artefacts and paralogues, a total of six putative SNP sequences were identified *in silico* that are predicted to map within the candidate trypanotolerance region on BTA7. Further work will be necessary to reveal whether the SNPs are indeed genuine.

4.4.3 Confirmation of comparative mapping predictions

The use of comparative mapping information to generate novel markers carries far less risk of contamination than chromosome microdissection. Furthermore, gene sequences amplified from cattle can be compared with existing sequences for the orthologous gene in other mammalian species to verify that the intended gene has been amplified, and that the product is not a human (or mouse) contaminant. For four of the five loci developed in this study, we observe clear homology between the database human and mouse sequences and the cattle sequence for the regions of exon DNA indicating that the sequence amplified is indeed the intended gene. At the same time, there are clear base differences indicating that the cattle product is not a human or mouse contamination artefact. For the fifth locus (FKBP38), the amplicon contains just 9bp of exon sequence, but this is conserved in human, mouse and cow. For all five loci, the intron amplified has very little homology to the human intron or any other database sequence, again suggesting that the sequence is genuine.

The gene sequences obtained show a high level of nucleotide polymorphism. Polymorphisms were detected in six of seven gene fragments sequenced. Overall, for three purebred animals representing the major bovine lineages (zebu, African taurine and European taurine), SNP frequency was one every 261bp. A comparable study of two genes (leptin and amyloid precursor protein, or APP) in a panel of zebu and European taurine cattle showed a mutation rate of one SNP every 89bp for leptin and one SNP every 19bp for APP (Konfortov *et al.* 1999). The increased SNP frequency is in part due to the larger number of animals (22) surveyed by Konfortov *et al.*. Considering nucleotide diversity instead, the level of diversity in the leptin gene ($2.6x10^{-3}$ nucleotide differences per site) is seen to be slightly higher than overall diversity for the seven loci sequenced here ($1.7x10^{-3}$). Diversity in the APP gene ($1.9x10^{-2}$) as reported by Konfortov *et al.* was an order of magnitude higher. SNP surveys in humans indicate significantly lower nucleotide diversities. A survey of approximately 2Mb of genomic and EST DNA in seven individuals found one SNP every 721bp, and an overall nucleotide diversity of $4.58x10^{-4}$, approximately 3.7 times lower. This probably reflects the fact that SNPs are more likely to be detected when studying highly diverse populations or individuals. The human study used European individuals, who probably have a recent common ancestor. In contrast, molecular clock estimates suggest that the evolutionary separation of the taurine and indicine (zebu) sub-species is of the order of hundreds of thousands of years (Loftus *et al.* 1994; MacHugh *et al.* 1997), allowing considerable time for mutation.

The divergence between zebu and taurine sub-species is considerably more ancient than both the divergence between African and European taurine cattle (estimated at 22,000 to 26,000 years BP) and the expansion dates for both taurine cattle lineages (of the order of 10,000 years BP) (Bradley *et al.* 1996). SNPs are therefore more likely to be detected when studying both sub-species of cattle, and of greatest use in studying zebu-taurine hybrid

animals. The discovery of SNPs at loci MANB and RAB3A, both previously reported to be monomorphic (Gu *et al.* 1999) supports this view. Moreover, for nine of the 18 SNPs detected in the three pure-bred animals sequenced here, the zebu animal is homozygous for one allele and the two taurine animals homozygous for the other, suggesting that these sites may be monomorphic in pure-bred animals from either sub-species. Additionally, we find pairwise nucleotide diversity between zebu and taurine sequences ($2.5x10^{-3}$) to be an order of magnitude higher than between the European and African taurine sequences ($2.1x10^{-4}$). This contrasts somewhat with the results of Konfortov *et al.* (1999), who found intra-taurine diversity to be 33% and 79% of zebu-taurine diversity for the two genes they studied, even though the taurine cattle surveyed were all European.

We had expected that SNPs would be more common in intron sequence, due to the reduced selection pressure against mutation. In line with this expectation, Hughes *et al.* (1997, 1998), reported a three-fold lower substitution rate for non-synonymous exon nucleotide sites than for intron sites for 42 genes in mice and rats. In human, a study of a 9.7kb region of the human lipoprotein lipase gene similarly showed intron polymorphism to be significantly more frequent (Nickerson *et al.* 1998). However, in this study, two SNPs were detected in a total of 388 bp of exon sequence (one SNP every 194bp), whereas 16 SNPs were found in a total of 4311 bp of intron sequence (one SNP every 269bp). Intron polymorphism was more frequent in one of two genes surveyed in cattle Konfortov *et al.* Polymorphism in the APP gene surveyed in cattle by Konfortov *et al.* (1999), was more than ten times as frequent in intron sequence than in exon sequence although for the leptin gene, polymorphism in intron and exon was approximately equally frequent. The results obtained so far in cattle may prove to be anomalous. There is no reason to expect that exon mutations should be relatively more frequent than in other species. We note that both exon polymorphisms reported here, and five of the seven reported by Konfortov *et al.* are at

synonymous sites, in accordance with expectation. We therefore believe that searching for SNPs in intron sequence will be more profitable.

The SNPs developed do not show clear discrimination between the N'Dama and Boran grandparent animals used to construct the trypanotolerance mapping pedigree at ILRI. Instead, both SNP alleles are seen in both pedigree founder breeds. It is possible that the SNPs do not in fact have genuinely zebu or taurine diagnostic alleles but that they are polymorphic in both zebu and taurine cattle. However, a more likely explanation is the high level of zebu / taurine genetic admixture in the pedigree founder animals. The Gambian N'Dama came from a region where there is substantial zebu genetic introgression into the native taurine breeds. Gambian N'Dama have an estimated 9-19% zebu nuclear genetic component (MacHugh *et al.* 1997). Kenyan Boran appear to be even more admixed. They are morphologically intermediate between zebu and taurine cattle, and show a high level of zebu / taurine genetic admixture (Frisch *et al.* 1997; Hanotte *et al.* In preparation). Considering only those SNPs that are homozygous for one base in the pure zebu typed and a different base in the pure taurine animals (**Table 4.7**), 50% of the alleles in the Boran founders are apparently taurine in origin, whereas only 4% of the N'Dama alleles are zebu in origin, supporting the view that Boran are the more admixed breed. Apparent recombination events are seen in three of four inter-genic regions in the four Boran founder animals. Of eight haplotypes, one appears to be a parental zebu haplotype. The remaining seven are recombinant, with six different haplotypes for a region spanning approximately 20cM (**Figure 4.5**). This is indicative of relatively ancient admixture with substantial subsequent recombination.

The poor discrimination between founder animals seen using the SNPs developed may make these markers less informative in trait-mapping studies. We would expect SNPs to

offer the greatest trait-mapping potential in pedigrees where the founder breeds were fixed for different alleles. Nevertheless, it would be interesting to see if there are differences in SNP allele frequencies for the F2 animals that are highly trypanotolerant and those that are highly trypanosusceptible.

4.4.4 Mapping genes in cattle

Gene mapping in cattle has been greatly facilitated by the development of somatic cell hybrid and radiation hybrid mapping techniques, and many genes originally mapped in human or mouse have now been mapped in cattle (Lyons *et al.* 1997; O'Brien *et al.* 1993). The method exploits the conservation of exon sequences between mammals, which allows primers designed using a human or mouse exon sequence to be used to amplify the orthologous cattle sequence. A potential complication, however, is that the PCR products obtained from the bovine and the hamster template in the cell hybrid may be indistinguishable on an agarose gel. Designing primers to amplify a stretch of DNA extending from one exon across an intron into an adjacent exon can circumvent this problem. Although intron positions in cattle are rarely known, the very high level of intron conservation between human and cattle (97% conserved) means that it is highly likely that an intron in human will also be present in cattle (K. H. Wolfe, personal communication). Introns in hamster and cattle are often of different sizes, allowing detection of the bovine PCR product. For SNP locus CALR, the hamster product is approximately 300 to 400bp larger than the cattle product (**Figure 4.3**) If the PCR products are of the same size, once the cattle intron sequence is known, cattle-specific primers may be designed (Gao and Womack 1997a; Gu *et al.* 1999; Gu *et al.* 1997; Kang'a *et al.* 2000). This was necessary for three of the SNP loci developed in this study (AKAP95, FKBP38 and JAK3).

4.4.5 Mapping confidence

The radiation hybrid map presented here has a very high level of statistical support (25 loci out of 33 ordered at Lod >=3.0). It is worth considering the factors contributing to this high level of support, as this may prove useful for future bovine radiation hybrid mapping studies.

4.4.6 Radiation hybrid panel information content

Locus retention frequencies observed in both panels in this study are relatively low when compared with other radiation hybrid studies. A review of RH mapping in human showed that, for 24 independent studies for which locus retention data is available, maximum locus retention frequency varied between 21% and 100%, with an average maximum retention frequency of 56% (Leach and O'Connell 1995). This compares with maximum retention frequencies for the BTA7 loci of 18.9% and 19.4% for the bovine panels. For the BovR5 panel, average locus retention frequency across all 30 chromosomes is 22.5%, with a low of 13.3% for BTA9 and a high of 45.3% for BTA19 (Band *et al.* 2000) (BTA19 contains the thymidine kinase gene, which was selected for in construction of the panel (Womack *et al.* 1997)). For the BTA7 loci typed in this study, average retention frequency in panel BovR5 is 13.3%. Similarly, average retention frequency in BovR12 for the BTA7 loci is low (13.1%), particularly when compared with the figure of 30.6% reported by Rexroad *et al.* (2000) for 18 loci on bovine chromosome 1 typed in 88 of the BovR12 hybrids. Both panels show appreciably lower retention frequencies than the optimal 50% required for maximal information content (Lange and Boehnke 1992).

The majority of the BTA7 loci can be successfully ordered (Odds >=1,000:1) using the BovR12 data alone: four of six group 1 loci; eight of fourteen group 2 loci, three of ten

group 3 loci and all three group 4 loci. These results are in accordance with predictions from simulations (Lunetta and Boehnke 1994). Lunetta and Boehnke carried out simulations with varying panel size, retention frequency and locus spacing to determine the number of loci that could be ordered at odds of 1,000:1. They noted that increases in retention frequency, RH panel size and marker density all increase the probability that a given number of loci can be successfully ordered. Results from simulations most closely corresponding to the BovR12 datasets for linkage groups 1 to 4 are shown in **Table 4.13**. Comparisons are not exact, as the retention frequencies for the BTA7 linkage groups (0.11 to 0.15) are lower than the retention frequency used in the simulations (0.2), and the number of hybrids in the BovR12 panel (160) is intermediate between the two values used in the simulations (100 and 200). Nevertheless, the simulation results suggest that, for each linkage group, the number of loci in the observed framework map agrees well with the number of loci we could expect to order. For linkage groups 1, 3 and 4, the simulations suggest that the probability of obtaining a framework map including the observed number of framework loci is close to or equal to one. The theoretical probability of obtaining the observed framework maps for group 2 is also good (p~0.36 to ~0.87).

Observed BovR12 panel data				Simulated data				
					100 hybrids		200 hybrids	
Linkage Group	No. Loci	No. Ordered	Distance (cR_{12000})	Distance (cR)	p(all)	p(obs)	p(all)	p(obs)
Group 1	6	4	179	170	~0.85	~0.97	~1	~1
Group 2	14	8	465	510	~0.03	~0.36	~0.33	~0.87
Group 3	10	3	208	170	~0.25	~1	~0.90	~1
Group 4	3	3	73	170	~1		~1	

Table 4.13: Comparison of linkage group framework maps obtained using BTA7 BovR12 data with simulation studies

Simulation data taken from Lunetta and Boehnke (1994). Locus retention frequency for simulations was 0.2. p(all) is the probability of ordering n_{tot} loci, where n_{tot} is the number of loci observed in the corresponding BovR12 linkage group. p(obs) is the probability of ordering n_{obs} loci, where n_{obs} is the number of loci successfully ordered (Odds of >=1,000:1) in the corresponding BovR12 linkage group.

The ability to place BTA7 loci into framework maps for each linkage group using the BovR12 data is in contrast with the results of a mapping study of BTA1 (Rexroad *et al.* 2000). Rexroad *et al.* typed 88 hybrids from the original BovR12 panel (subsequently expanded to 160 hybrids) for 18 loci from the centromeric region of BTA1. Although the number of hybrids used is half that used in this study, the average locus retention frequency (30.6%) was more than double that for the BTA7 loci (13.1%). Superficially, it might therefore appear that the two datasets should have comparable potential for resolving locus order. However it was not possible to obtain a statistically ordered map using the BovR12 BTA1 data.

The key difference between the BTA1 and BTA7 BovR12 datasets that accounts for the difference in mapping potential appears to be the increased locus density for the BTA7 loci. Simulations have shown that locus density affects the probability of obtaining an accurate locus order (Lunetta and Boehnke 1994). Lunetta and Boehnke determined the probability of ordering 3 equally spaced markers with varying inter-marker distances (cR). For panels of varying hybrid number (20 - 100) with retention frequency 0.5, the

probability of identifying the true order as the most likely is highest for inter-marker distances between approximately 20 and 70 cR (breakage probability $\theta \approx 0.2$ to 0.5). For locus separations below 20 cR, probability of obtaining the correct order falls sharply. Other simulations using 100 hybrids with retention frequency 0.3 also showed that θ values from 0.2 to 0.5 give the best chance of determining the true order, while for $\theta = 0.8$ (~160 cR), the probability of obtaining the correct order is substantially reduced (Barrett 1992). For the framework loci identified using the BovR12 BTA7 data alone, mean locus separation is 47.1 ± 10.9 cR_{12000}, and breakage probability θ is 0.33 ± 0.05 (for 14 pairwise distances between loci in the same linkage group). For all loci (non-framework loci included), mean locus separation is 32.8 ± 4.1 cR_{12000} ($\theta = 0.26 \pm 0.03$). Both figures are in the optimal range indicated by simulation.

It is not possible to compare physical locus separation in centi-rads between the BTA1 and BTA7 BovR12 datasets, as this data is not available for the BTA1 loci. However, it is possible to compare genetic linkage distances. Using distances from Kappes *et al.* (1997), average separation for loci in the BTA1 RH map of Rexroad *et al.* is at least 3.0cM. The separation may be higher, as not all of the loci typed by Rexroad *et al.* are included in the linkage map. In contrast, average separation for loci in our BTA7 RH map is approximately 1.7cM (Gu *et al.* 2000). The relatively small locus separation for the BTA7 loci accounts for the strong linkage observed between them. All 33 loci typed in the BovR12 panel were linked at Lod 3.0, whereas for the 18 BTA1 loci, seven distinct linkage groups were obtained at Lod 3.0.

Another factor contributing to the high confidence of ordering loci using the BovR12 data is the relatively even locus spacing. Lunetta and Boehnke (1994) demonstrated that, for three loci, the probability of obtaining the correct order is greatest if the central marker is

equidistant from the first and third markers in the order, and declines exponentially as the central marker moves closer to either of the flanking markers. Nevertheless, even when the distance 'd_1' between the closest two markers is just 10 percent of the total distance 'd_T', the probability of obtaining the correct locus order is still approximately 65 percent of the probability when d_1 equals $d_T/2$. Simulations by Jones (1996) also showed that equidistant loci are ordered with higher confidence (Jones 1996). Within each linkage group, the loci are all relatively evenly spaced (BovR12 data). For 25 sets of three loci, d_1/d_T averages 0.33 ± 0.02.

The map constructed using just the BovR5 panel data contains only four framework loci compared with 18 for BovR12 (**Table 4.11**). The relatively poor resolving power of the BovR5 panel is likely to be due to the lower number of hybrids and the relatively close spacing between loci in each linkage group (20.8 ± 3.5 cR_{5000}). This is a consequence of the low number of breaks occurring along the chromosome (**Table 4.10**) due to the low dose of radiation used. Other mapping studies using panel BovR5 have reported higher percentages of framework loci, but the overall retention frequency has been higher in all these studies (Band *et al.* 1998; Ozawa *et al.* 2000; Rexroad *et al.* 1999; Schlapfer *et al.* 1997).

The use of the two mapping panels together contributes greatly to the high statistical confidence of the final map. The map contains more framework loci than either of the maps constructed using just panel BovR5 or panel BovR12 (**Table 4.11**). Similar findings have been reported in human mapping studies. A map of human chromosome 21 constructed using two panels together contained more framework loci than maps constructed using either panel alone (Lunetta *et al.* 1996). Another study demonstrated that two whole-genome maps constructed using separate panels could be integrated to give a

map that was more accurate than either (Agarwala *et al.* 2000). The enhanced mapping power observed is at least in part due to the increased amount of information available on pooling datasets. Additionally, though, it is possible that panels of different resolution provide complementary information. Panels with a low level of chromosome breakage may be better at revealing long-range order whereas panels with a high level of breakage may be better at resolving local locus order. In this study and the two studies mentioned (Agarwala *et al.* 2000; Lunetta *et al.* 1996), there is a significant difference in the resolution of the two panels used.

4.4.7 Comparison of RH map with BTA7 genetic linkage map

Comparing the radiation hybrid map obtained with the comprehensive genetic linkage map (Gu *et al.* 2000) reveals two differences in the locus order for the 12 loci common to both maps (**Figure 4.5**). The more significant difference is in the position of the gene LDLR. The genetic linkage map places LDLR after microsatellite BM6105, which would locate it near to the end of the RH map whereas here it is found to be located in the centre of the map. The second difference is that the positions of the adjacent microsatellite loci BP41 and BL1067 is interchanged. In both cases, it appears that the RH map order is the more likely. Both loci are framework loci in the RH map (ordered at odds of 1,000:1) whereas neither is a framework locus in the genetic linkage map.

In addition to locus order we may also compare inter-locus distances between the maps. Ignoring locus LDLR, which appears to be incorrectly positioned in the linkage map, distance between adjacent loci correlates well for loci RM012 to BM6105 (Pearson 'r' coefficient 0.94, 6 d.f., $p<0.01$), although it appears that there is a compression of genetic distance between BMS713 and BL1067 / BP41. From RM012 to BM6105, 1cM on the genetic linkage map corresponds to approximately 44 cR_{12000}. In contrast, the next region

of the map appears to show a significant expansion of genetic distance relative to physical (RH) distance. Between loci BM6105 and BM7247, a distance of approximately 20cM on the linkage map corresponds to a distance of approximately 125 cR_{12000}, making 1cM equal to 6.3 cR_{12000}. There are various explanations for discrepancies between radiation hybrid and genetic linkage maps. The rate of recombination is known to vary considerably along a chromosome, so that genetic distance increases relative to physical distance in regions of frequent recombination, and decreases where recombination is rare (Lichten and Goldman 1995). It has also been suggested that the probability of a break occurring during irradiation may also vary along the chromosome (Raeymaekers *et al.* 1995).

4.4.8 Comparison of RH map with published bovine radiation hybrid maps

Two radiation hybrid maps of BTA7 have been published (Band *et al.* 2000; Gu *et al.* 1999). These maps were both produced using the same 5000 rad panel used in this study. The first map of Gu *et al.* (1999) has seven loci in common with our map. These loci are, in order from the centromeric end, RM012, COMP, ILSTS001, MANB, IL4, BM6105 and UWCA20. Of these, four loci are framework loci: RM012, ILSTS001, BM6105 and UWCA20. All seven loci occur in the same order in our map and that of Gu *et al.*

The second map, that of Band *et al.* (2000), has eight loci in common with ours (**Figure 4.7**). In order from the centromere these loci are COMP, RM012, ICAM3, BM6105, RM006, IL3, IL4 and UWCA20. Of these, COMP, RM006, IL4 and UWCA20 are framework loci. There are three rearrangements of locus order between the two maps. Near to the centromere, loci COMP and RM012 are inverted. It is not possible to determine which order is the more likely as only one of the two loci is a framework locus in both cases. The second rearrangement is in the position of locus BM6105, the seventh of the eight common loci in our map, whereas in that of Band *et al.* it is the fourth. In this case,

the position of BM6105 in our map has higher confidence, suggesting that this is the more likely position. The third rearrangement is that the neighbouring loci IL3 and IL4 are transposed. Here it is not possible to determine which orientation is more likely, as only one of the loci is a framework locus in both cases.

Although inter-locus distances are not published for the two radiation hybrid maps, it generally appears that, for loci whose position in the order is the same as found here, the distances are comparable.

4.4.9 Comparison of RH map with human physical map

For comparing the bovine RH map to the human physical map, we have used the sequence map of HSA19 produced by the LLNL Genome Center (Ashworth *et al.* 1995), available at http://greengenes.llnl.gov/genome/genome.html. BLAST® searches using the 'Golden Path' HSA19 assembled contig sequence (7th Oct 2000 dataset, obtainable from http://genome.cse.ucsc.edu) were conducted to find positions of genes included in the bovine radiation hybrid maps and not on the LLNL map. The positions of the genes on the LLNL map were first determined, and the positions of the unmapped genes relative to these genes were then found. The resulting map is shown in **Figure 4.7** (centre). Comparison confirms that the centromeric region of BTA7 corresponds essentially to the p arm of HSA19, with gene order broadly conserved. The LLNL map places the HSA19 centromere at approximately 22-24Mb. It therefore appears possible that the BTA7 centromere corresponds to the HSA19 centromere, although the map of Band *et al.* (2000) shows a single HSA1 gene at the centromeric end of the map. Our map shows an unbroken region of conserved synteny between HSA19 and BTA7 extending for approximately 10Mb. Distal to this region on BTA7, there is a cluster of cytokine genes (IL4 to CSF2) mapping close together on HSA5 (within 10 cR_{3000} in human (Deloukas *et al.* 1998)). The

map published by Band *et al.* suggests that this gene cluster interrupts the region of conserved synteny between HSA19 and BTA7. However, there is only a single HSA19 gene (GADD45B) distal to the two HSA5 genes on their BTA7 map, and there is only weak support for this orientation. It is therefore at present unclear whether the region of conserved synteny between HSA19 and BTA7 continues beyond the cytokine cluster.

Within the pattern of conserved order between HSA19 and BTA7, there are a number of apparent rearrangements. Three inversions are required to account for differences between our map and the HSA19 map (**Figure 4.7**). First, loci JAK3, FKBP38 and COMP at the centromeric end of our map are inverted relative to the human map. As is noted above, the bovine radiation hybrid map of Band *et al.* (2000) shows an inversion of the centromeric end relative to our map. If the order proposed by Band *et al.* is correct, this might remove the need to invoke an inversion between the human and cattle maps for this region. A second inversion is needed to account for the reversed orientation of loci DNMT1, ICAM3, LDLR and EPOR. However, if this were the case, we would also expect the positions of loci DNM2 and ICAM1/ICAM3 to be inverted in the map of Band *et al.*, which is not seen. The locus order is more strongly supported, and more likely to be correct. A final inversion between our map and the human map occurs for loci MYO1F, ARP(NG27) and NDUFA7. Two additional rearrangements involving locus OAZ1 and DNASE2/ETR101 would be necessary to account for differences between the map of Band *et al.* and the human map. In both cases, however, support for the order proposed by the Band *et al.* is weak, so it is unclear if the rearrangements are genuine.

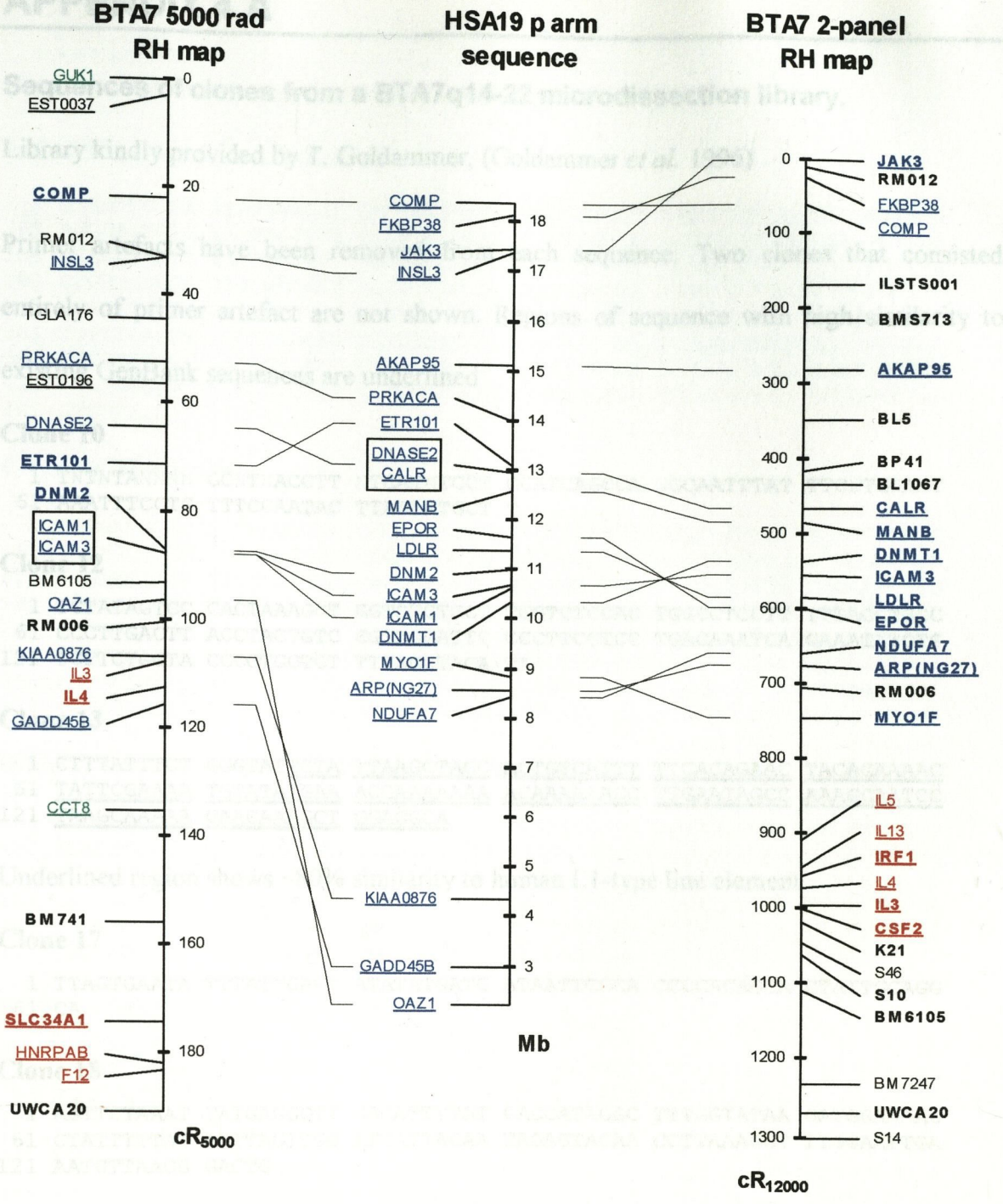

Figure 4.7: Comparison of radiation hybrid maps of the centromeric region of bovine chromosome 7 with human chromosome 19 p arm

Left: bovine radiation hybrid map of Gu *et al.* (1999). Centre: human chromosome 19 p arm sequence map. Right: two-panel bovine radiation hybrid map.

Loci in bold type are framework loci, ordered at odds of ≥1,000:1. Underlined loci are genes (Type I loci). Genes in blue type map to HSA19; genes in red type map to HSA5; genes in green map to other human chromosomes. Boxes enclose loci mapping to the same location. Where available, positions of HSA19 genes (in Mb) are taken from the LLNL map of HSA19 (Ashworth *et al.* 1995). For genes not included in the LLNL map, positions relative to LLNL markers were determined by BLAST® searching using the Golden Path HSA19 contig sequence (7th Oct 2000 dataset) obtainable from http://genome.cse.ucsc.edu.

APPENDIX 4.A

Sequences of clones from a BTA7q14-22 microdissection library.

Library kindly provided by T. Goldammer, (Goldammer *et al.* 1996)

Primer artefacts have been removed from each sequence. Two clones that consisted entirely of primer artefact are not shown. Regions of sequence with high similarity to existing GenBank sequences are underlined

Clone 10

```
  1 TNTNTANNNN GCNTNACCTT NTCATNTCCT CCATCAGCCA NGCAATTTAT TTCNTGNANT
 61 AAATTTCCTG TTTCCAATAC TTAGAATGCT T
```

Clone 12

```
  1 TCTATAGTCC CACTAAAGCT GGTCTCTCCT TGGTCTCCAC TGGCCTCCTT TTTACCTGCC
 61 CCCTTGACTT ACCTACTGTC CGTAGGATTC CCCTTCCTCC TGACAAATCA GAAATATGTG
121 TCTTCTGGTA CCCCTGCTCT TTTAGCTACA TA
```

Clone 13

```
  1 CTTTATTTCT GGGTATTCTA TTAAGCTACC ACTGTCATTT TTCACAGAAC TACAGAAAAC
 61 TATTCGAAAA TGTATATGAA ACCAAAAAAA ACAAAAAAGC TTGAATAGCC AAAGCAATCC
121 TAAGCAAAAA GAACAAAGCT GGAGGCA
```

Underlined region shows ~80% similarity to human L1-type line element

Clone 17

```
  1 TTAGTGAATA TTTATTGACT ATATATGATG ATAATTCCCA CCCCACAGCA CTATTGCAGG
 61 CA
```

Clone 18

```
  1 AGTTCTAAAT TATGAGGGTT GACATTTTAT GACCATAGGC TTTGGTATAA GATGGAAAAC
 61 CTATTTTTAT ATTTAGTTGG ATTATTACAA TAGAGTACAA CCTTAAAGCT TTGGAATTGA
121 AATGTTAACG GACTG
```

Clone 19

```
  1 AATTCTCCAG GCAAGAATAC TGGCGTGGGT TGCTGTTTCC TTCTCCAGGC ATCTTCCCGA
 61 CCCCTGGAAC GAAACCAGGT TTCCTGCATT GCAGGCAGAT TCTTTACCGT CCAAGCTACC
121 GGGGAAGCCC ATGGTTGTGT CCATGGTCAG TTCATAGGAT AATAAAATCG TGATGTCCCC
181 AATCACCTAG CAGTTTA
```

Underlined region shows extensive similarity (~87%) to Bos taurus SINE-type repeat

Clone23

```
  1 TATTCCTAAA TCTCAACTGA TGAATCTTTC TACCTGTAAT AATGTTGTTC CGTTAGTTCG
 61 TTTTATTAAC GTAGATTTTT CTTCCCAACG TCCTGACTGG TATAATGAGC CAGTTCTTAA
121 AATCGCATAA GGTAATTCAC AATGATTAAA GTTGAAATTA AACCATCTCA AGC
```

100% identity to phage M13 sequence

APPENDIX 4.B

Consensus sequences for seven loci mapped to BTA7

AKAP95 (A-kinase anchoring protein)

```
  1 GGAGAGGTGA TTATAGACCT GAGTCCTTCC CCGGGGTCCC ACGTGCCCCG CTTTGTATAG
 61 AGGGAGTGGG CACCTGCTCG TCCTGGGAGG GCTCTTCCCC AGCTCCCTCC TCAGTGTAGG
121 GCAGATAAGT TGAGCCTTTG TTGGGGCAGA GGCATTTGGT AGGACTTGCT GGTTTCCTGG
181 GTGGTGGGCC TTAGGGCATG CCCGTTAGCT CAACTATGTA AGACCCCTTA CTTCTGCATC
241 AGGCACCCAG CATCTTCTCC CTGTTCCTGy GGCCAGGGTT TCACTCCCCT TGTTCAGTGG
301 TGTTTCAGAG GGATTCAGTA TGTTTTACTT TGTGAGCATC TCAGAGGGGT GCATGGCAGT
361 TTGGTTTTGC TGCTCTGAAT GTAGGGAGGT GTGCTTTGCA GGAGGCGTTG AGATCGTGCC
421 TGGACAGTGT GGTGATGGGT TGTGCTTTCT TTGCAGCTCC TTTCGCTTCC AGTCGTTTGA
481 GTCCTATGAT TCC
```

CALR (Calreticulin)

```
  1 CATAATGTTT GGTGAGGGGA CCCCACCCTG GTGCTAATTT TTGGTCACTT AGAGGGAGTT
 61 TAAAACCCCG GGAGAGCCCA TGAACGGTGG TCTTGAGAGC ACAGAAArCA CTTAAATGCT
121 TTAGACATAT TTTCGATAGG rAAGGTGAGG GAATAGTTAs GAATGACTTA CTGCATTTTG
181 ATCATCGACy TAAGACTCGT GCAATTTTAC TATATAGGCA TTTCCTCGTT CCTAGTTGTG
241 CAAGTGAAGA TACTGAGTTC TGGAGCACTG ATTTTTTTTT TGAAGTGTCC TCCCCCAGTT
301 CCTTAGAGAG AAGGTCyGAA GTGTGTTTTT TCAGCTTGAC AGTCCTGAGT TAAAGAACCA
361 AGTTGTCTTT TTTTAATGGG GTGGGCAGGG GCTTGTAATG GTGGACAGAC TGTAGTTGTG
421 ACCAGATCTG CTTTTCATCA AGGCCCGGAC ATCTGTGGCC CTGGCACCAA
```

DNMT1 (DNA (Cytosine-5-)-methyltransferase 1)

```
  1 TTCAGCAAAG CCGATATATG AGGATGAyCC ATCTCCCGAA GGTAAGGGGA ATGTTCCAGA
 61 TTGTCTGAAA CAGTCAGTTT GTAAACCGAA CAGAGTCTGG GCTGyGAGCA GCTACAAGCA
121 GTCTyGCTCT GATGGCTCCA GTAmGGCTGC CCGGCTCACC TGATCAGAGG TGTGGGCTCG
181 GGTGGGGACT CAGTGCCACT GAACCTGCTT GGGTTTTTGA GGCATATTCT TTTAGTCATG
241 TCCAACTsTG CAACCCCATG GACTGTAGCC TGTCAAGGCT CCTCTGTCCA TGGGATTTTC
301 CAGGCAAGAA GACTGGAGTG AGTTGCCATT TCTTAGTCCA CTTCATTTCA AGAAACC
```

FKBP38 (FK506-binding protein 8)

```
  1 TCCAACAAGG TGAGCAGAAT GGGGGCACCC TGAGACTTGC CATGCTGACC TGTCGGTAAG
 61 TTCCTCTTGT GGTCTGCCCC TTATCTTTCT TGCTGTGGCT CTCCCCTATA GAGTTTCTGT
121 CTGGGCAGTA GCAGGGGGTA CATGACTGAA GGGCCCCAGC CTGCCTGCTG AGACCyTTTT
181 TGACCTCTGC AGAGCCAGTC TGGGCTCTCA CCCTCAGCAG GGGTTCCAGA GGTCAAGGGC
241 ACTGCCAGCC TGGCCAGTGT CAGTGTTTTC AGCTTCTTTC TCAGAATATT TCTTCTCGGA
301 AACCTTyTTC TGGCCACCCT TGCTCCACCT CATCTTCTGG CTTCCTCGGA GGATGCAGAC
361 AGCACCCCCC ACTTTGCTAA TGCTTGAAGA GTAGTAGACA GCTCCCACGT CCCAGTGGTT
421 CCAGCCTCAA yGAGCCCTGG ACCCCAGGGA GGCAGGCTGC TAGCAGCAGT GGTTGCCCCT
481 GCTC
```

JAK3 (Janus kinase 3)

```
  1 TGAGCCAAGT GTCATACCAG CATCTCGTAT TGCTCCACGG CGTGTGCATG GCTGGAGACA
 61 GTGAGAGACA CTCCCCATCT CTGCCCTGCC TCACTGGGCC CACCTACAGC AGCAATAGCA
121 GTACACAATC AGAGCCTGCC CTGGCTGGGG CTCCATTCCA CCAGGGTCTA TCATCTTGTA
181 CTCATGTACC TCCATCATCA CTCTTTGAAT AAGCAGCTCC TGGCTCCCCC CAATACTTAA
241 GCCCCCAGAA TTATCTCAAG TCAGGCAGAC CTGGGTTCTG ATCTATTGTG TACCATCTCT
301 ACAGCTTTGC AAGGTCCATG CTCCCTCTCC AAGCCTTGTT TCCGTCAGTT CAGCTCAGCA
361 GCTGTTTATT GGGAGCCCAA CATGTGCCAG GAGCTGGGTG CATAGCATAG AACATGACAG
421 GCTGAGCAGG ATGATAAATG ATTGGGTGTG GCAGGGCAGC ACAGGTCATT GTGGGTGATC
481 AGAGGCTGCT AACCCACTTG GAGGTCAAGC AGTGATTCTG AGTAGGTAAG GATACATTCT
541 AAGCTGGAAA GGCAAGAATA AATAGACATT AGGGAAGTGA GAGAAGAGGG GTGTTCCAGT
601 CAGAAACACA GAACTTGCAA AGGGTTGGGA GTGAGTCATG GAGAGAGTGC AACACCTTCT
661 AGGCATTACA GGCAAATGGT TCTGAGATGG GAAGTAGGGA AGCAGAGAGG TGAATAGGGT
721 AAGGATGGAG GAATAGGGGA GGGAAAAGTC AGAGCTGAGA GTTCAAGCGA CCCTTGAAGG
781 TGGGGTTGGG GGCAGAGCCT GTTGAATACA ATACAGTCCT ACCTCCCAAC CCTGC
```

MANB (Lysosomal alpha-mannosidase)

```
  1 ATACGCCATT TGCAGCCTCT ATGTGGCTGC CGTTGCTGCT kCTGCTAAGT CGCTTCAATC
 61 GTGTCCGACT CTGTGTGACC CCATAGACAG CAGCCCACGA GGCTCCCCCA TCCCTGGGAT
121 TCTCCAGGCA AGAACACTGG AGTGGATTGT GTTCCATAGC TCTCTTACAC TGGCCTGAGA
181 GTGACCCCTG ACCCTTCTCC TCTCArGCCT GGTCGTTTAG GCAGGTCTTC TCGTCCCTGG
241 CATCACCAAC CCTGGCGCCA CTCCTGGCCC TGACAACTGA CTTGGACTTT GCCCCTCCCG
301 GCACAGGACG CCTAGAGTTT GCCAACGGTG GCTGGGTGAT GAACGATGAG GCGACCACCC
361 ACTACGGAGC CATCATCGAC CAGATGACAC TCGGACTGCG CTTCCTGGAG GAGACGTTCG
421 GCAGCGACGG GCGCCCCCGT GTGGCyTGGC ACATCGACCC ATTCGGCCAC TCTCGGGAGC
481 AAGCTTCACT GTTCGCGCAG GTTTTCAGAT CTCTTGGGCC CGCCCCTTCA TTCCTTCTGA
541 CTCCTCCTCT GTCATCCAAG CCCCGCCCTT TTCTGCAAGT TCACCCGAAC CCGAACCAGG
601 CCCTACCCCT GGCCCTCTCG CCACTTAAGA CCCTGCCTCT TGGGTGACCT GTGAATCCCA
661 TTCTTTTTGC TCTGGCCTTG GCTCTGCTCT GTCCTAGCCT AGGTTGACCC TCATCACCTA
721 TTCCCATACA CCCCGGCCTC CCTTGTCCAG CTGAGTCTTC CCCCTCCCCT GATTCCATCC
781 AGTTTGCTCT GATCCTGGCT CTTGGCCAAG CTGGGTGGCT GGTTAGGGCC TGCGGTTTTT
841 CAGCCAGCTT CGTTACCTGT GCCATGACCA TCCCCCTTCC CTTGCTAGGC TGCGGAAAGG
901 GGAAGGCCTC CTAGAGCTGG GGAAGGTGGA GGTGGGTTGT AGCGTTGGAC TTGCTCCCTC
961 CCGGCGCTTC CGCAGGGCTT CTGACCTTCC TCAGCCTTTG AAATGAACTG GAGGGCCTCG
1021 CTGGTCTTTG ACTTGGTTTT TCTCCCTGTG CGTGAGAGGT TGGTGGTTGG TGATGAGAGG
1081 ACCGGTCCCT TATGCATCCT GCCCTCTCTT GTTCTyCCAT CCyACTCGTC ATCCCTCAA
```

RAB3A (RAS oncogene family member)

```
  1 TTACACACTG GGCATTTATG AGGCGCTTAC TGTGTACCAG GGGCTAGAGG CACAGAAGAG
 61 AGTGTCCCCT TTCTCCCATC AGTGCTCTCC AGGGGTGGGG AAGGGGTAAA TGCTGGGTCC
121 TGTCCTACCC CTGACATCAT CAGATGCATT GCTTGCTTAG TGCCTCAGTT TCCCCATGTT
181 CAAGTTCCTC CAyGAGTGTC TGGATGAGGT TTCAGAGGGA TCCTGACCCC TCAACCTCAT
241 TCCCTGGGCC AAGAGCAGGC TCCTCCACCA GGGTCTTCTG GCCACCTCCC TCAACACCTC
301 TGGGTCCTAA AGGGATCTGA TAAATCCCAA GCCAGAGCTG GTTGTCCCCT GCTCACAATC
361 CTTCCATGGC TCCCACCTCA CTCTTTTTTT TTTTTCCTTT TTTTTTTTTC CACCTCACTC
421 TTGGTAACAC CTGCAACCTT CCCCTGGCTG AGGAGACCCT GCCTGGTCAG CCCCTGTTGG
481 TGCCTCATTT CATCTGTCTC CCTCTGTCTC CAGCCTCTTC ATTGTTCCTC AAATGTACCA
541 GACCCAGTCC CACTTCAGGG CATTTGCCTA GACTTyCCTT CCCCTGACCC CCGTATGACT
601 CCCTCTCTCA TCCTGCAAGT TTTTAAATTG TCCCCCATCC CACACTGCCC ATCCTTGCCT
661 GGTTTTTTTT CTTCTCCTGT CTCATTACTT TCTGACAGTA CATTTCCGTG ATCTGAGGAT
721 GCATTTTCTT TCTTGTCCTG GGTTGATTCT GTTTCCTGCA CAGAGCTGTC AGCTCCAAAA
781 GGGCAGTGAT TTGGGCCCAT TTTGTTAACG ACTGTGTCCT CAGTGCCCAG CACAGGGTGT
841 GGCACAGGTC CCTATGTGTC GAATGAAGAA GGACTCTCCC CCAACCA
```

APPENDIX 4.C

Putative centromeric BTA7 SNP loci from GenBank sequences

SNP_Seq1

```
  1 GGGTGGGCGA CGTGCACCCA GAGTATTAAC TTCCCGAGAG GTCAGCGTGC CAATATCCCA
 61 GGGCAGGCCT GGCCCCGCAG TGCAAAACCT CCCCATCCGG CCCACTGGCA CCCCACCAAC
121 TTGGCCCCAC CTCyGCGCCG GGAGAATGGG GTGGCAGGTT TCAGGCCCTC CTCGCCCCCA
181 GTATGAACTT CACCTCTGCG GCGCATCCTT GAGTGCCAGC CCCCACCCCC ATGTCACATC
241 ACCCAGCAGA ACTGCCCTCC GAGTGCCAAC GCCCCAGTGG CGTGAGGAGC CCTGCAGCCG
301 GCGGGGGCAA CTTCACCTGC TTGTATATTA AACAATCCTG ATTTCAGACA ATTTAAATCT
361 TAATCTATTT AAAAAAGAAT ATTATATAAA GATGCTGTTT TTAAACCTTT TGTCATTTGA
421 GATGCATGTA TCTTCCGTGC GGGCTGAGGG GCGGGTGGTC AGCTCAGTGC CGGCCTCTGG
481 GGGCTGCTGG AGCCGCCCAT GTTCAGGCAG TGCGTCCGCA TTGGGGCCCA CACCCCTTAA
541 TCCAGAGAAA GCTGGTGGAA ATTGACTTTT CAT
```

SNP_Seq2

```
  1 GAACTCCGTA GTCCAGCTGC GGCTGGAAGG TCCGAGGCTG TCTGCCGCCT TCGGTGCCTG
 61 TTTTGGAGGA TGGATCCGGG GCCCGGGGCC AACGGGATGC CGTTGGCTGG CTTAGCCTGG
121 TCGTCGGCCT CGGCGCCCCC GCCGCGGGGA TTCAGTGCGA TCTCCTGCAC TGTGGAGGGG
181 ACGCCCGCCA GCTTCGGCAA GACTTTCGCT CAGAAATCTG GCTACTTCCT GTGCCTCAAT
241 CCTTTGGGCA GCCTAGAGAA TCCACAGGAG AACGTGGTrG TCGACATCCA GATCCTAGTG
301 GACAAGAGCC CCCTCCCGCC GGGATTCTCA CCAGTCTGCG ACCCCCTGGA CTCGAAGGCC
361 TCCGTGTCCA AGAAGAAACG CATGTGCGTG AAGCTGGTGC CCTTGGGGGC CGCGGACACA
421 GCTGTCTTTG ACATCCGACT GAGTGGGAAA ACCAAGACAG TCCCTGGATA CCTGCGAGTA
481 GGGGACATGG GGGGCTTTGC CATTTGGTGC CGGAAGGCCA AGGCCCCTCG GCCAGTGCCC
541 AAGCCCCGAG CTCTTAGCCG AGACGTGAGG GACCTCTCCC TGGACTCGCC GGGCCAG
```

SNP_Seq3

```
  1 CGTGGGGATT TTGTAATGCC GACCGGAGAC TTTGATTCGA AGCCCAGTTG GGCGGACCAG
 61 GTGGAAGAGG AAGGAGAGGA CGACAAATGT GTCACCAGCG AGCTCCTCAA GGGGATCCCC
121 CTGGCCACTG GGGATACCAG TCCAGAGCCT GAGCTACTGC CGGGAGCTCC ACTrCCGCCT
181 CCCAAGGAGG TCATCAATGG AAACATCAAG ACAGTGACGG AGTATAAGAT AGATGAGGAT
241 GGCAAGAAGT TCAAGATTGT CCGCACCTTC AGAATTGAGA CCCGGAAGGC CTCAAAGGCT
301 GTGGCAAGGA GGAAGAACTG GAAGAAGTTT GGGAACTCAG AATTTGACCC ACCGGGGCCC
361 AACGTAGCTA CCACCACAGT CAGCGATGAT GTATCCATGA CATTCATCAC CAGCAAAGAG
421 GATCTGAACT GCCAGGAAGA GGAGGATCCA ATGAACAAGC TCAAGGGCCA GAAGATAGTG
481 TCCTGCCGAA TCTGCAAGGG CGACCACTGG ACCACCCGCT GCCCCTACAA GGACACGCTG
541 GGGCCCATGC AGAAGGAGCT GGCCGAACAG CTGGGCCTGT CCACTGGCGA GAAGGAGAAG
601 CTCCCCGGAG AGCTGGAGCC TGTGCAGGCC ACTCAAAACA AGACTGGGAA GTAC
```

SNP_Seq4

```
  1 GTGACCCGGA TGATGGCGGC CGGCACGGCC TTAGGCTTGG CCCTGTGGCT ACTACTGCCG
 61 CCAGTGGGCG TGGGAGGGGC AGGGCCCCCG CCGATCCAGG ACGGCGAGTT CACGTTCCTG
121 CTGCCTGCGG GGCGGAAGCA GTGTTTCTAT CAGTCCGCGC CGGCCAACGC AAGCCTyGAG
181 ACTGAGTACC AGGTGATCGG AGGTGCTGGG CTGGACGTGG ATTTCAGTCT GGAGAGCCCT
241 CAGGGAGTGC TGCTGGTCAG CGAGTCCCGC AAGGCAGACG GTGTGCACAC GGTGGAGCCC
301 ACGGAGGCCG GGGACTACAA GCTGTGCTTT GACAACTCCT TCAGCACCAT CTCGGAGAAG
361 CTGGTGTTCT TTGAACTCAT CTTTGACAGC CTGCAGGATG AGGAGGAGGT CGAGGGCTGG
421 GCAGAGGCTG TGGAGCCTGA GGAGATTCTG GAAGTCAAGA TGGAGGACAT CAAGGAGTCC
481 ATCGAGACCA TGAGGATCCG GCTGGAGCGC AGCATCCAAG TGCTGACTCT GCTGCGAGCC
541 TTTGAGGCAC GTGACCGCAA CCTGCAAGAA GGCAACCTGG G
```

SNP_Seq5

```
  1 GACAAAACGA AGTAATTCTG ATGATGGACT GGATTCTGCA GGCCACTTCC CACCATCATG
 61 GCCTCCATCT CCAGGCCAGC CTTGCTAGAG TCTGCCTCCA GTTCTGTAGG CTTAGGGGTT
121 GTGGGAACTG GGGCTTTGAT GTATGAGTAG AAGTTCACCA AATATAGACA GAGGAGTCAT
181 GCACTTCACA GGCAGACACT GATGCCCCAG GAGGCAGAGG CCTCAAAGGG CACTGCCAGG
241 GTAGGGGAAG AATGTsCACC CTGGACAAGA TTGTTTTGTA TTTGATGCCT ACCAGGGTGA
301 GAACTTGACC ATGATGTGTC TAAAACACCC CCTGGGAAGT GCCCTCGATA TGTCAGAGCC
361 TTGCTGGGAT GAAATCCGAA TAAGAAAATC CATCTGGGGG TCAGTGAGGC CCAAGTTGCA
421 ATCTTGGCTT CACTGGCTCT GGGAAAATGG CTTCCCCAAC TCTGTGCCTC AGTTTCCTTG
481 TGTTTACAAA ACTAATACTA ATGACTATTT ATTAAGCACC TACTGTATGC CAAGCGCTTT
541 TACTTGTGGC TTCTCCCTCA GTCAGTCT
```

SNP_Seq6

```
  1 GGCTCCTCCG GCTCCTGGTC CGGCTTCTGG CGGCTCCGGG GAGGTGGACG AGCTGTTCGA
 61 CGTGAAGAAC GCCTTCTACA TTGGCAGCTA CCAGCAGTGC ATCAACGAGG CGCAGCGGGT
121 GAAGCCATCC AGCCCGGAGA GAGATGTGGA GCGGGATGTC TTCCTGTACA GAGCATACCT
181 GGCCCAGAGG AAGTACGGCG TGGTGCTGGA CGAGATCAAG CCCTCCTCCG CCCCGGAGCT
241 GCAGGCCGTG CGCATGTTTG CTGAGTACCT GGCCAGCGAC AGCCGGCGGG ATGCGATCGT
301 GGCCGAGCTG GACCGrGAGA TGAGCCGGAG CGTGGATGTG ACCAACACCA CCTTCCTGCT
361 CATGGCTGCC TCCATCTATT TCTACGACCA GAACCCAGAT GCAGCCCTGC GCACCCTTCA
421 CCAGGGGGAC AGCCTGGAGT GCATGGCCAT GACAGTGCAG ATCCTGCTGA AGCTTGACCG
481 CCTGGACCTT GCCCGGAAGG AGCTGAAGAA GATGCA
```

CHAPTER 5

CONCLUSIONS

5.1 Effects of selection at genes on linked microsatellites

The results described in **Chapter 2** show that microsatellites can be influenced by selection at linked genes. Observations from real populations are found to agree with theoretical predictions concerning the effects of different modes of selection on allelic diversity and the shape of allele distributions. Overdominant selection is known to act at MHC class II genes and, as predicted, nearby microsatellite loci show higher allelic diversity and more even allele distributions when compared with loci randomly distributed throughout the genome. Again in line with theory, inter-population genetic distance between distantly related populations is also found to be lower at MHC-linked microsatellites, presumably due to a reduction in genetic drift. For microsatellite loci linked to genes, the effects are essentially opposite in nature. Allelic diversity is reduced and allele distributions skewed, as expected from theory. One anomalous finding is that the rate of increase of inter-population genetic distance appears to be slower at gene-linked microsatellites than for loci not linked to any selected loci. More work will be needed to investigate why this might be so. One unknown factor is how far along the chromosome the influence of selection at a gene extends. The MHC-linked microsatellites span a distance of 2.6cM in the linkage map of Kappes *et al.* (1997) and all show apparent selective effects. It is possible though that there are many loci throughout the 2.6cM stretch that are subject to overdominant selection, so the maximum distance of any of the MHC-linked loci to a selected locus is unknown. The gene-linked loci are all very close (typically within several kb) of the nearest gene. Evidence that the effects of selection tend to be restricted in range is that the majority of microsatellites are apparently not influenced by selection, even though it is likely that most will be reasonably close to genes.

5.2 Lack of selective influence at microsatellites near trypanotolerance QTL

Application of the analytical methods used in **Chapter 2** to microsatellites from three QTL regions for trypanotolerance in African cattle in **Chapter 3** finds that there is no evidence that the loci have been influenced by selection at nearby genes responsible for trypanotolerance. This may be because the selective effects are too short-range to influence microsatellites several cM away. An additional complication is that there is substantial inherent variation between microsatellite loci that may mask the effects of selection, particularly if the number of loci surveyed is small. Increasing the number and density of markers studied in wild populations may yet reveal that certain microsatellites near the QTL studied have indeed been affected by selection.

Another approach to detecting loci selected for trypanotolerance may be to investigate hybrid populations. If selection has been very intense, the percentage genetic component of the selected, taurine breeds to hybrid populations may increase with increasing proximity to the trypanotolerance genes. Such an approach has been used successfully in mapping the locus responsible for warfarin selection in rats (Kohn *et al.* 2000). The same approach did not prove informative when attempted here, using microsatellites to search for an excess of taurine alleles in two hybrid populations. A complicating factor is that individual microsatellite loci differ greatly in their potential to discriminate between the ancestral zebu and taurine populations. Using loci that discriminate clearly between zebu and taurine cattle would overcome this problem. SNP loci could fulfil this role. It is therefore desirable to develop a panel of SNP loci for which one allele is fixed in taurine cattle and a different allele in zebu cattle, and which map to the QTL regions.

5.3 Developing polymorphic markers within targeted chromosomal regions

Results from **Chapter 4** demonstrate that information from comparative mapping studies between human and cattle can be used reliably to predict which genes are likely to be located within a particular bovine chromosomal region. The evident high level of conservation of gene structure between human and cattle means that the position of introns within these genes in cattle can also be predicted. Conservation of flanking exon sequence in human and other species can then be exploited to design primers to amplify across the predicted bovine intron. This is important because intron sequences are much more divergent between species than the corresponding exons, hence cattle intron sequences can be clearly distinguished from those of other species. This provides confirmation that the amplified sequences are genuine and not the result of contamination with human or other DNA. More importantly, it means that primers can be designed to amplify exclusively in cattle so that the loci can be amplified in bovine/hamster radiation hybrid mapping panels. This provides further data for refining the bovine-human comparative map. We note that only mapping of genes can achieve this. Anonymous SNP loci contained within bovine 'junk' DNA do not contribute to inter-species comparative maps.

A second reason for targeting bovine introns for amplification is that is likely that they will be more polymorphic than exons, due to lower selective pressure. Introns are therefore more likely to yield SNPs for population genetic studies and genetic linkage mapping. This is not seen for the introns amplified in this study, but this may be an artefactual result due to the small number of introns amplified. In the small number of cattle tested, we observe a high level of sequence variation (1 SNP per 261bp). Most of the SNPs detected distinguish between zebu and taurine animals, reflecting the ancient date of sub-species divergence. A high proportion of bovine introns can thus be expected to yield SNPs distinguishing zebu and taurine cattle.

A shortcut to SNP detection may be provided by the wealth of bovine sequence data in the public domain. To date, there are almost 200,000 cattle DNA sequences in GenBank, of which over 90% EST sequences obtained from cDNA libraries. The number of EST sequences far exceeds the number of bovine genes, hence there are many duplicated sequences as demonstrated in **Section 4.3.14**. If duplicated sequences were obtained using different EST libraries prepared from different individual cattle, we can expect them to show nucleotide variation. Confirmation of this is provided by the detection of six apparent bovine SNP loci with human homologues mapping to the region of HSA19 corresponding to the BTA7 trypanotolerance QTL region.

The strategies discussed above have the potential to provide novel SNP loci that map to a specific chromosomal region and that discriminate clearly between zebu and taurine cattle. Such loci could be genotyped both in the ILRI trypanotolerance mapping pedigree to provide further information on QTL location. Additionally they could be typed in hybrid West African field populations, as discussed, to investigate how hybrid composition varies in proximity to QTL for trypanotolerance.

CHAPTER 6

SOFTWARE DEVELOPED

6.1 THE EXCEL MICROSATELLITE TOOLKIT

The Excel Microsatellite Toolkit is a Visual Basic® for Applications (VBA) utility for Microsoft® Excel that provides a number of tools for working with haploid or diploid microsatellite data. The utility can be used with Microsoft® Excel 1997 or later versions for Windows®.

6.1.1 Data Checking

The tools highlight as invalid any alleles that are non-numeric, non-integer, negative or outside a user-defined range. Suspiciously large gaps in allele size are also reported. For diploid data, samples with only one allele are highlighted. Duplicate and invalid sample or population names are also reported.

6.1.2 Data Formatting

Data can be formatted for a range of commonly used population genetics programs.

Formats available:

Arlequin (Microsatellite, Standard and Frequency formats) (Schneider *et al.* 1997)

Microsat (population-level and individual-level formats) (Minch 1995)

GenePop (Raymond and Rousset 1995) (also read by Genetix (Belkhir *et al.* 2001) and Bottleneck (Cornuet and Luikart 1996))

Fstat (Goudet 1995)

Dispan (Ota 1993)

Formatted data can be saved directly from Excel for reading by the relevant program.

6.1.3 Calculations

Allele frequencies can be calculated and matching samples can be detected. If the number of samples is below 255, a matrix of allele sharing coefficients can be outputted. The coefficients are equal to the proportion of alleles shared between a pair of individuals. Bowcock *et al.* (1994) have denoted this measure *Ps*. They go on to derive a distance $1 - Ps$ for the construction of phylogenetic trees with individuals as operational taxonomic units (Bowcock *et al.* 1994). A matrix of $1 - Ps$ values can be created using the Excel microsatellite toolkit. This can be saved and read by programs in the Phylip package (Felsenstein 2001). Other statistics calculated are observed and expected heterozygosity and mean number of alleles per locus.

6.1.4 Download and use

The toolkit can be downloaded as a zip file 'MStools3.zip' over the Internet from http://acer.gen.tcd.ie/~sdepark/ms-toolkit/

The zip file contains 4 files:

MS_tools.xla	Excel add-in containing VBA objects and code
MS_data.xls	Sample data for use with MS_tools.xla
MS_tools.hlp	Help file for MS_tools.xla
Readme.txt	Installation information

Due to the object-oriented nature of VBA, with its reliance on forms created and edited on-screen, there is no stand-alone program code that it would be appropriate to list here. The code and objects contained within the application are, however, freely accessible and may be viewed and edited by any user wishing to do so.

6.2 SNP HUNTER

SNP Hunter is a Perl application created to facilitate the process of identifying single nucleotide polymorphisms in sequences obtained from the GenBank database.

SNP Hunter requires an installation of the Perl environment (obtainable from http://www.perl.com) to work. Although Perl is supported on Unix®, Windows® PC and Apple® Macintosh® operating systems, several features of the SNP Hunter code (concerning screen clearing and file locations) are specific to Windows® PC. These could easily be altered so that the code works on other operating systems.

In its current form, SNP Hunter for PC assumes that three different directories are used for downloading files from NCBI, running BLAST® homology searches, and running ClustalW sequence alignments. The program has default locations for these directories, although they can be reset to any valid directory path.

Outline of method for finding SNPs for a given species:

- Download all sequences for a given species.
- Use BLAST® program (Altschul *et al.* 1997) to compare each sequence with all others and identify pairs of sequences with a very high degree of homology.
- Identify non-redundant groups of homologues from BLAST® results.
- Produce a sequence alignment for each group of homologues.
- Scan alignments to identify polymorphic sites.

6.2.1 Procedure for identifying SNPs in GenBank sequences

Steps are listed in the order in which they would typically be performed. Not all steps are necessarily required. Steps performed by SNP Hunter are listed as Option A, Option B etc. Other steps are marked with bullet points.

- Download sequences for a species using 'batch entrez' from NCBI (http://www.ncbi.nlm.nih.gov/Entrez/batch.html)

  If there are fewer than 20,000 sequences in the database, it is possible to download all sequences in FASTA format. Otherwise, it is necessary to download the GI numbers (unique identifiers to DNA and protein sequences) for the species first, and then to split them into batches for downloading.

Option A of SNP Hunter allows you to split a list of GI numbers into batches of 20,000 for downloading using batch entrez.

Option B extracts a list of GI numbers from a file of sequences in FASTA format. In thus way it is possible to obtain a list of sequences that have already been downloaded.

Option C compares two lists of GI numbers, outputting only GI numbers exclusive to one or other list. It is thus possible to compare a list of GI numbers of sequences that have already been downloaded with the updated list of GI numbers from GenBank, and output only the new GI numbers to a list for downloading. This removes the need to download sequences that have previously been downloaded.

Option D assembles multiple files of sequences in FASTA format into a single file

Option E is used to detect unrecognised database identifiers present immediately before accession numbers in the definition lines of sequences in FASTA format. GenBank accession numbers are preceded by a code indicating the database to which the sequence was originally submitted (e.g. gb = GenBank, dbj = DDBJ, emb = EMBL).

The program searches for this code when trying to identify GenBank accession numbers in BLAST® results files in order to reduce the BLAST® results file to a manageable tabular format (**Option I**). Unrecognised codes can lead to accession numbers not being read, hence the formatted BLAST® results may be incomplete. This is overcome by using option E to return any sequence definition lines for which accession numbers cannot be determined due to unrecognised database identifiers. The program settings can then be updated using **Option S** to allow for recognition of new database identifiers.

Option F enables removal of sequences above or below specified lengths. This is useful if only EST (Express Sequence Tag) sequences are desired. Long genomic sequences may span multiple groups of ESTs, potentially causing problems in alignment. Removal of sequences over 1000 bp - only about 2% - should help

- Create a BLAST® database from the FASTA format sequence file using 'formatdb' (obtainable from ftp://ncbi.nlm.nih.gov/blast/executables)
- BLAST® the database against itself using the stand-alone version of BLAST®. Use a high expectation ('E') value to minimise false positives (<=10e-100). Turning off sequence alignment output will reduce the size of the output file. Specify a reasonable number of hits (homologues found at the given expectation value for each sequence tested) to output (default is 500).

Option G reads a file listing sequence GI or accession numbers and can then either extract the sequences from a batch of FASTA sequences or return all sequences from the batch except those whose GI / accession numbers are listed.

Option H can be used to resume if BLAST® crashes or is stopped part way through. The option reads through the BLAST® results file to find how many sequences were

successfully analysed, and constructs a new BLAST® input file which excludes these sequences.

Option I formats the results from the BLAST® output file into a simple tabular output. It is possible at this stage to specify required confidence level for BLAST® hits. Hits below this level are discarded.

Option J combines multiple BLAST® output files either before or after formatting using Option I

Option K identifies groups of homologues in the formatted BLAST® results file. Within each group, every sequence has significant homology to at least one other group member. No member of any group has significant homology to sequences in any other groups. The minimum number of sequences required in a group can be specified. Groups with fewer sequences than this number are rejected.

Option L finds the longest sequence in every group and outputs to a file; it also extracts the sequences belonging to groups of homologues from the original database to a new FASTA format file to create a new database for subsequent BLAST® treatment.

- Use Formatdb to create a new database from the file containing just the sequences that belong to homologue groups.

- BLAST® the longest sequence from each homologue group against all sequences in the new database at relatively low stringency, specifying to output a large number of hit sequences to pick up every member of the homology group (1000 or more), and also to output sequence alignments, and hence relative strand orientations.

Option M extracts sequences for each homology group from the BLAST® database and outputs them in the correct orientation to a file with a .seq extension correct for

alignment using ClustalW. It is possible to specify minimum and maximum group sizes at this point. Groups with fewer than the minimum number of sequences are rejected. Groups with more than the maximum number of sequences are reduced in size by removal of sequences at random - although the longest sequence is always retained.

Option N creates a file (seqfiles.txt) listing all of the .seq files in the ClustalW directory. This is done automatically under Option M, but if .seq files were subsequently deleted or added to the ClustalW directory, it is necessary to recompile the list before proceeding to the sequence alignment stage.

Option O performs quick alignments of the sequences with ClustalW for each .seq file in turn. This stage is best carried out using multiple instances of the SNP Hunter program, aligning a separate batch of 50 files in each instance. Attempting to get ClustalW simply to align all .seq files in the directory usually fails as ClustalW stops responding after processing about 50 files, hence the need to create a separate file listing all of the .seq files before starting on the alignments and then breaking the list into batches of 50 files for alignment.

Option P reads through each alignment (.aln file) and identifies polymorphic bases. It is possible to specify either a minimum absolute number of copies of the rare allele (>=2), or a minimum frequency of the rare allele (<=50%). It is also possible to specify how many bases at the start and end of each sequence should be ignored to reduce the problem of reporting of false polymorphism due to artefactual sequencing errors. Option P also outputs a consensus sequence for each group of homologues.

- As a final stage, the BLAST® program can be used to search for similarity between the consensus sequences and sequences from individual human chromosomes. Information on the chromosomal rearrangements between human and the target species can then be

used to identify SNPs mapping to a region corresponding to a specific chromosomal fragment in the target species. If the consensus sequence shows similarity to multiple human sequences with different genomic locations, this may be taken as an indication that the homologous sequences found in the target species are in fact paralogous members of a gene family. Conversely, if a consensus sequence shows similarity to only one human sequence, this suggests that the homologous sequences found in the target species are genuine.

APPENDIX 6.A: SNP HUNTER LISTING

```
#SNP_Hunter.pl - written by Stephen Park, Dept of Genetics, TCD, Feb 2001
use Win32::Console;
use Cwd;
my $current_dir = cwd;
my $cons = Win32::Console->new(STD_OUTPUT_HANDLE);

&initialise;
while ($option !~ /^X$/i) {
    $option = 0;
    while ($option !~ /^([a-pA-PsSxX]{1})$/) {
        $cons->Cls;
        print "\tSNP-HUNTING SOFTWARE  -  PLEASE SELECT AN OPTION:\n\n";
        for $option (sort {$a cmp $b} keys %options) {
            print "  $option\t$options{$option}\n";
            if ($option =~/[GLPX]/){print "\n";}
        }
        print "  Option: ";
        chomp ($option = <>);
        $option = uc ($option);
    }
    exit 0 if $option eq "X";
    $cons->Cls;
    print "\nOption $option: $options{$option}\n";
    if ($option eq "A") {&split_gi_list;}
    if ($option eq "B") {&get_db_gi_list;}
    if ($option eq "C") {&cut_old_gis;}
    if ($option eq "D") {&combine_seqfiles;}
    if ($option eq "E") {&new_db_identifiers;}
    if ($option eq "F") {&cut_long_short_seqs;}
    if ($option eq "G") {&accs_or_gis_from_seqfile;}
    if ($option eq "H") {&rescue_blast;}
    if ($option eq "I") {&blast_results;}
    if ($option eq "J") {&combine_blast_results;}
    if ($option eq "K") {&blast_groups;}
    if ($option eq "L") {&extract_groups;}
    if ($option eq "M") {&create_seqfiles;}
    if ($option eq "N") {&clustal_filelist;}
    if ($option eq "O") {&batch_clustal;}
    if ($option eq "P") {&get_snps;}
    if ($option eq "S") {&prog_setup;}
    &close_filehandles;
}

##########################################################################

sub initialise {
    if (-e "SNP_Hunter_Setup.txt"){
        return 0 if ! &open_file ("","SETUP","input","SNP_Hunter_Setup.txt",$working_dirs{P});
        while (<SETUP>){chomp; eval;}
        close (SETUP);
        }
    else {
        #The next line contains sequence database identifiers. Newly-encountered database identifiers must be
added.
        (@db_ids) = split (/\s/, 'emb gb dbj ref bbs pdb');
        #The next line contains the working directory for ClustalW
        $clustal_dir = 'C:\GenProgs\Clustal';
        #The next line contains the working directory for BLAST
```

```
    $blast_dir = 'C:\Blast'; #Updatable
    #The next line contains the working directory for NCBI files
    $ncbi_dir = 'C:\Windows\Desktop\NCBI'; #Updatable
}

$invalid_dirs = 0; $new_ncbi_dir = 0;
$cons->Cls;
print "\nSNP_HUNTER: CHECKING WORKING DIRECTORIES\n\n";
if (! -e $blast_dir || ! -d $blast_dir) {
    print "Invalid working directory for BLAST: $blast_dir\n";
    print "You must specify a valid directory for BLAST\n\n";
    $invalid_dirs++;
}
if (! -e $clustal_dir || ! -d $clustal_dir) {
    print "Invalid working directory for ClustalW: $clustal_dir\n";
    print "You must specify a valid directory for ClustalW\n\n";
    $invalid_dirs++;
}
if (! -e $ncbi_dir || ! -d $ncbi_dir) {
    print "Invalid working directory for NCBI files: $ncbi_dir\n";
    print "You must specify a valid directory for NCBI downloads\n\n";
    $invalid_dirs++;
}
$working_dirs{P}=cwd;
$working_dirs{C}=$clustal_dir;
$working_dirs{B}=$blast_dir;
$working_dirs{N}=$ncbi_dir;

if ($invalid_dirs) {print "Use the 'setup' option from the menu. Press Return to continue.";<>;}
elsif($new_ncbi_dir) {print "Press Return to continue.";<>;}

$options{A} = "Split GI list from NCBI into blocks of 20,000 GIs for downloading";
$options{B} = "Get GI list from existing sequence file (or database)";
$options{C} = "Remove old GIs from new GI list to download new GIs only";
$options{D} = "Combine batches of 20,000 sequences into a single file";
$options{E} = "Find unrecognised sequence database identifiers in sequence def-lines";
$options{F} = "Remove long / short sequences from batch file";
$options{G} = "Output / remove batch of accessions / GIs from batch sequence file";
$options{H} = "Resume after BLAST crashed / stopped before completion";
$options{I} = "Fillet BLAST results";
$options{J} = "Combine multiple BLAST results files";
$options{K} = "Find groups of duplicate sequences in BLAST results";
$options{L} = "Extract groups from database and longest sequence from each group";
$options{M} = "Create input files for ClustalW alignment for all groups";
$options{N} = "Get list of .seq files in ClustalW directory for alignment ";
$options{O} = "Launch ClustalW for aligning sequences";
$options{P} = "Find SNPs in ClustalW alignment files";
$options{S} = "Change program setup";
$options{X} = "Exit program (press Ctrl-C to exit at any time)";
}

##########################################################################

sub done {my $sleep_time = @_; print "\nFinished.\n"; sleep $sleep_time;}

##########################################################################

sub close_filehandles {
    my $fh;
    for $fh (keys %open_filehandles){
        close $fh;
    }
```

```
    undef %open_filehandles;
}

#########################################################################

sub split_gi_list {
    my ($outfile_stem, $longoutname); my $j=1;
    return 0 if ! &open_file ("GI list to break up","GI_LIST", "input");
    print "\nEnter the stem for the OUTPUT FILE name:\n";
    print "(Output files will take the form of 'stem'_1.out, 'stem'_2.out etc)\n";
    chomp ($outfile_stem = <>);
    return 0 if $outfile_stem =~ /^$/;
    $outfile_stem =~ s/\.\w+//;
    OUTER: while (<GI_LIST>) {
        if ($. % 20000 == 0) {
            close(OUTFILE);
            $longoutname = $outfile_stem . '_' . $j++ . "\.out";
            return 0 if ! &open_file ("","OUTFILE","output",$longoutname,cwd);
            print "Outputting GIs to file $longoutname ...\n";
        }
        print OUTFILE;
    }
    &done(1);
    close (GI_LIST);
    close (OUTFILE);
}

#########################################################################

sub get_db_gi_list {
    return 0 if ! &open_file ("sequence / database file","SEQFILE", "input");
    return 0 if ! &open_file ("GI list","OUTFILE", "output");
    print "\nOutputting GIs...\n";
    while (<SEQFILE>) {
        if(m/(?<=^>gi\|)\d{1,}(?=\|)/) {print OUTFILE "$&\n";}
    }
    &done(1);
    close (SEQFILE);
    close (OUTFILE);
}

#########################################################################

sub cut_old_gis {
    my %old_gis;
    return 0 if ! &open_file ("*OLD* GI LIST","OLD_GI_LIST", "input");
    return 0 if ! &open_file ("*NEW* GI LIST","NEW_GI_LIST", "input");
    return 0 if ! &open_file ("GIs exclusive to new list","OUTFILE", "output");
    print "\nThinking...\n";
    while (<OLD_GI_LIST>) {
        chomp;
        $old_gis{$_} = 1;
        }
    close (OLD_GI_LIST);
    while (<NEW_GI_LIST>) {
        chomp;
        next if exists ($old_gis{$_});
        print OUTFILE "$_\n";
        }
    close (NEW_GI_LIST);
    close (OUTFILE);
```

```
    &done(1);
}

#########################################################################

sub combine_seqfiles {
    my %seq_files; my $file = 0; my $msg = ""; my $dir_choice = 0;
    while ($dir_choice !~ /^[NCB]$/i) {
        $cons->Cls;
        print "\nOption $option: $options{$option}\n";
        print "\nSelect directory containing sequence files:\n\n";
        print "  (B) - BLAST:    $working_dirs{B}\n";
        print "  (C) - ClustalW: $working_dirs{C}\n";
        print "  (N) - NCBI:     $working_dirs{N}\n\n";
        chomp ($dir_choice = <>);
        return 0 if $dir_choice=~/^$/;
        $dir_choice = uc $dir_choice;
    }
    chdir ($working_dirs{$dir_choice });
    while ($file !~ /^$/){
        $cons->Cls;
        print "In directory ".cwd."\n";
        print "\nOption $option: $options{$option}\n";
        print $msg;
        print "\nEnter name of sequence file " . (scalar (keys %seq_files) + 1);
        print " (just hit return when finished):\n";
        chomp ($file = <>);
        last if $file =~ /^$/;
        $file = lc $file; #Files case-insensitive in MS-DOS
        if (! -e $file){$msg = "\nNo file \'$file\' in directory " .cwd. "\nPlease try again\n";}
        elsif (exists ($seq_files{$file})) {
            $msg = "\nYou\'ve already entered file \'$file\'. Please try again.\n";
        }
        else {$msg = "\nFile \'$file\' OK\n"; $seq_files{$file}=1;}
    }
    return 0 if scalar (keys %seq_files) == 0;
    if (scalar (keys %seq_files) == 1) {print "You can\'t combine just one file!\n";sleep 1;return 0};
    return 0 if ! &open_file ("combined sequence files","OUTFILE", "output");

    print "\nCombining files...\n";
    for $file (keys %seq_files) {
        return 0 if ! &open_file ("","INFILE","input",$file,cwd);
        while (<INFILE>) {
            print OUTFILE;
        }
        close (INFILE);
    }
    close (OUTFILE);
    &done(1);
}

#########################################################################

sub new_db_identifiers {
    my $unknown_db_ids = 0;
    return 0 if ! &open_file ("FASTA sequence / database file","SEQFILE", "input");
    return 0 if ! &open_file ("unrecognised database identifiers","OUTFILE", "output");
    print "\nSearching for unrecognised database identifiers...\n";
    OUTER: while (<SEQFILE>) {
        if(m/^>/) {
            for $db_id (@db_ids) {
                if ( m/(?<=$db_id \|)[^\|]{1,}(?=\|)/ ) {next OUTER;}
```

```
            elsif ( m/(?<=$ db_id \|\|)[^\s]{1,}(?=\s)/ ) {next OUTER;}
            elsif ( m/(?<=$ db_id \|).{1,20}(?=\s)/ ) {next OUTER;}
        }
        $unknown_db_ids ++;
        print OUTFILE;
    }
}
close (SEQFILE);
close (OUTFILE);
if (! $unknown_db_ids) {print "\nNo unrecognised database identifiers found\n";}
else {print "\n$unknown_db_ids definition lines found - see output file for details.\n";}
&done(4);
}

#########################################################################

sub cut_long_short_seqs {
    undef $min_len, undef $max_len; undef $seq_def; undef $seq; undef $seq_count; undef $err;
    $min_len = 0; $max_len = 0; $seq_count = 0; $err = 0;
    return 0 if ! &open_file ("FASTA sequence / database file","SEQFILE", "input");
    return 0 if ! &open_file ("sequences","OUTFILE", "output");
    while () {
        $cons->Cls;
        print "\nOption $option: $options{$option}\n";
        if ($err){print $err;}
        print "\nEnter the minimum sequence length required:\n";
        chomp ($min_len = <>);
        return 0 if $min_len =~ /^$/;
        print "\nEnter the maximum sequence length required:\n";
        chomp ($max_len = <>);
        return 0 if $max_len =~ /^$/;
        if ($min_len =~ /\D/ || $max_len =~ /\D/){$err = "\nERROR: Min and Max lengths must be
numbers\n";next}
        if ($min_len < 1 || $max_len < 1){
            $err = "\nERROR: Min and max length must be at least one base pair\n"; next;
        }
        if ($min_len > $max_len){$err = "\nERROR: Max length cannot be less than min length\n"; next;}
        last;
    }
    if ($min_len==0) {$min_len=1;}
    print "\nOutputting sequences between $min_len and $max_len bp...\n";
    while (<SEQFILE>) {
        chomp;
        if(/^>/) {
            &print_seq_if_ok;
            $seq_def = $_;
            $seq = "";
        }
        else {$seq .= $_;}
    }
    &print_seq_if_ok;
    close (SEQFILE);
    close (OUTFILE);
    print "$seq_count sequences between $min_len bp and $max_len bp";
    undef $min_len, undef $max_len; undef $seq_def; undef $seq; undef $seq_count; undef $err;
    &done(4);

    ###################################

    sub print_seq_if_ok{
        if (length $seq >= $min_len && length $seq <= $max_len) {
            $seq_count++;
```

```
            print OUTFILE ">$seq_def\n";
            while (length $seq > 0) {
                print OUTFILE substr($seq,0,70,"")."\n";
            }
        }
    }
}

#########################################################################

sub accs_or_gis_from_seqfile{
    my (%accs, $acc, $output_seq, $seq_count);
    my $gi_list = 1; my $acc_list = 1; my $output_option=0; my $response =0;
    return 0 if ! &open_file ("FASTA sequence / database file","SEQFILE", "input");
    return 0 if ! &open_file ("list of accessions / GIs","ACCESSIONS", "input");
    print "\nReading input...\n";
    while(<ACCESSIONS>) {
        chomp ;
        $accs{$_} = 1 ;
        if (/^[^A-Z]/){$acc_list = 0;}
        if (/\D/){$gi_list = 0;}
    }
    close (ACCESSIONS);
    if ($gi_list && ! $acc_list){print "\nData read - type is list of GIs\n";sleep 2;}
    elsif ($acc_list && ! $gi_list){print "\nData read - type is list of Accession Nos.\n"; sleep 2}
    else {print "\nCannot determine if input list contains GIs or Accessions\n";
        while ($response !~ /^(y|yes|n|no)$/i){
            print "\nDo you want to treat the list as a list of Accessions anyway? (y/n)";
            chomp ($response = <>);
            return 0 if $response =~ /^(|no?)$/;
        }
        $acc_list = 1;
    }
    return 0 if ! &open_file ("sequences","OUTFILE", "output");
    while ($output_option !~ /^[12]$/) {
        $cons->Cls;
        print "\nPlease select an option:\n";
        if ($gi_list) {
            print "\n1: Only output sequences for GIs in GI list file\n";
            print "\n2: Only output sequences where GI is not in GI list file\n";
        }
        else {
            print "\n1: Only output sequences for Accession Nos in list file\n";
            print "\n2: Only output sequences where Accession No is not in list file\n";
        }
        chomp ($output_option = <>);
    }
    print "\nOutputting sequences...\n";
    $seq_count=0;
    if ($output_option == 1){
        while (<SEQFILE>) {
            if(/^>/) {
                if ($acc_list){$acc = &get_name($_);}
                elsif (/(?<=^>gi\|)\d{1,}(?=\|)/){$acc = $&;}
                else {$acc = "";}
                if (exists ($accs{$acc})){
                    delete $accs{$acc};
                    $output_seq = 1;
                    $seq_count++;
                    }
                else {$output_seq = 0;}
                }
```

```
            next if ! $output_seq;
            print OUTFILE;
        }
    }
    elsif ($output_option == 2){
        while (<SEQFILE>) {
            if(/^>/) {
                if ($acc_list){$acc = &get_name($_);}
                elsif (/(?<=^>gi\|)\d{1,}(?=\|)/){$acc = $&;}
                else {$acc = "";}
                if (exists($accs{$acc})) {$output_seq = 0;}
                else {$output_seq = 1; $seq_count++;}
                }
            next if ! $output_seq;
            print OUTFILE;
        }
    }
    close (SEQFILE);
    close (OUTFILE);
    print "Outputted $seq_count sequences";
    &done(4);
}

################################################################################

sub rescue_blast {
    my $query_count = 0; my $seq_count = 0;
    return 0 if ! &open_file ("BLAST results file","BLAST", "input");
    print "Counting sequences successfully BLASTed...\n";
    while (<BLAST>) {
        if(/^Query=/) {
            $query_count++;
        }
    }
    close (BLAST);
    $query_count--;
    print "$query_count sequences were successfully BLASTed.\n";
    print "A new BLAST input file will now be created by cutting these sequences\n";
    print "from the start of the old BLAST input file.\n";
    return 0 if ! &open_file ("old BLAST input file","INFILE", "input");
    return 0 if ! &open_file ("remaining sequences to BLAST","OUTFILE", "output");
    print "Creating new BLAST input file...\n";
    while (<INFILE>) {
        if (/^>/) {$seq_count++};
        next if $seq_count <= $query_count;
        print OUTFILE;
    }
    close(INFILE);
    close(OUTFILE);
    &done(2);
}
################################################################################

sub find_qry_gaps {
    return 0 if ! &open_file ("BLAST results file","BLAST", "input");
    return 0 if ! &open_file ("query sequence file","QUERYSEQ", "input");
    return 0 if ! &open_file ("gapped query sequence","OUTFILE", "output");
    while (<QUERYSEQ>){
        chomp;
        if (/^>/) {$qry_name = $_;}
        else {$qry_seq .= $_;}
    }
```

```
    (@qry_seq) = (split (//, $qry_seq));
    close (QUERYSEQ);
    while (<BLAST>) {
        chomp;
        if (m/^Query: /) {
            $qry_line = $_;
            $qry_line =~ /(?<=^Query: )\d+/ ;
            $aln_start = $& ;
            $qry_line =~ /(?<=\s)\d+$/ ;
            $aln_end = $& ;
            $qry_seq = $qry_line ;
            $qry_seq =~ /(?<=\s)[^\s\d]+(?=\s)/;
            $qry_seq = $&;
            while ($qry_seq =~ /\-+/g){
                $len = length $&;
                $qry_to_gap = $` ;
                $qry_to_gap =~ s/\-// ;
                    # $locn is the position in the original query sequence of the gap
                $locn = $aln_start + length($qry_to_gap) - 1 ;
                    # the position of the gap is entered in two hashes - one for the original query sequence
                    # and the second for the blast hit
                if (exists($gaps{$locn}) && $gaps{$locn} < $len){$gaps{$locn} = $len;}
                else {$gaps{$locn} = $len;}
            }
        }
    }

    for $locn (sort {$a<=>$b} keys %gaps){
        $qry_seq[$locn-1] = $qry_seq[$locn-1] . "n" x $gaps{$locn};
    }

    close (BLAST);

    $qry_seq = join '',(@qry_seq);
    print OUTFILE "$qry_name\n";
    print OUTFILE $qry_seq;
    close(OUTFILE);
}

################################################################################

sub extract_blast_alignment {
    return 0 if ! &open_file ("BLAST results file","BLAST", "input");
    return 0 if ! &open_file ("query sequence file","QUERYSEQ", "input");
    return 0 if ! &open_file ("gapped query sequence","OUTFILE", "output");
    while (<QUERYSEQ>){
        chomp;
        if (/^>/) {$qry_name = $_;}
        else {$qry_seq .= $_;}
    }
    $qry_len = length $qry_seq;
    (@qry_seq) = (split (//, $qry_seq));
    close (QUERYSEQ);
    while (<BLAST>) {
        chomp;
        if (m/^Query: /) {
            $qry_line = $_;
            $qry_line =~ /(?<=^Query: )\d+/ ;
            $aln_start = $& ;
            $qry_line =~ /(?<=\s)\d+$/ ;
            $aln_end = $& ;
            $qry_seq = $qry_line ;
```

```
        $qry_seq =~ /(?<=\s)[^\s\d]+(?=\s)/;
        $qry_seq = $&;
        while ($qry_seq =~ /\-+/g){
            $len = length $&;
            $qry_to_gap = $` ;
            $qry_to_gap =~ s/\-// ;
                # $locn is the position in the original query sequence of the gap
            $locn = $aln_start + length($qry_to_gap) - 1 ;
                # the position of the gap in the original query sequence is entered in a hash
            if (exists($gaps{$locn}) && $gaps{$locn} < $len){$gaps{$locn} = $len;}
            else {$gaps{$locn} = $len;}
        }
    }
}
for $locn (sort {$a<=>$b} keys %gaps){
    $qry_seq[$locn-1] = $qry_seq[$locn-1] . '-' x $gaps{$locn};
}
seek BLAST, 0, 0;
while (<BLAST>) {
    chomp;
    if (m/^>/) {
        $hit_id = &get_name($_);
        $hits{$hit_id} = 0 ;
        if (! exists ($aln{$hit_id})) {@{$aln{$hit_id}} = ('.') x $qry_len;}
    }
    elsif (m/^Query: /) {
        $qry_line = $_;
        $qry_line =~ /(?<=^Query: )\d+/ ;
        $aln_start = $& ;
        $qry_line =~ /(?<=\s)\d+$/ ;
        $aln_end = $& ;
        $temp_qry_seq = $qry_line ;
        $temp_qry_seq =~ /(?<=\s)[^\s\d]+(?=\s)/;
        $temp_qry_seq = $&;
        <BLAST>;
        $temp_sbj_seq = <BLAST>;
        $temp_sbj_seq =~ /(?<=\s)[^\s\d]+(?=\s)/ ;
        $temp_sbj_seq = $&;
        (@temp_qry_seq) = (split (//, $temp_qry_seq));
        (@temp_sbj_seq) = (split (//, $temp_sbj_seq));
        $#temp_qry_seq = -1;
        for ($i = 0; $i <= $#temp_qry_seq; $i++){
            if ($temp_qry_seq[$i] eq '-'){
                $temp_qry_seq[$i-1] = $temp_qry_seq[$i-1] . $temp_qry_seq[$i];
                splice (@temp_qry_seq, $i, 1);
                $temp_sbj_seq[$i-1] = $temp_sbj_seq[$i-1] . $temp_sbj_seq[$i];
                splice (@temp_sbj_seq, $i, 1);
                $i --;
            }
        }
        for ($i = 0; $i <= $#temp_qry_seq; $i++){
            if (length($qry_seq[$aln_start-1 + $i]) > length($temp_qry_seq[$i])) {
                $aln{$hit_id}[$aln_start-1 + $i] = $temp_sbj_seq[$i] . '-';
            }
            else {
                $aln{$hit_id}[$aln_start-1 + $i] = $temp_sbj_seq[$i];
            }
        }

    }
```

```
    }
    for $blast_hit (keys %aln){
        $seq = join ",(@{$aln{$blast_hit}});
        print OUTFILE "$blast_hit\t$seq\n";
    }
    close(OUTFILE);
}

seek BLAST, 0, 0;

#############################################################################

sub blast_results {
    undef $alignments; undef $query; undef $hit; undef $score; undef $expect; undef @hitline;
    print "\nExtracts query and hit names, scores and (where applicable)";
    print "\nstrand orientations from BLAST files\n";sleep 2;
    return 0 if ! &open_file ("BLAST results file","BLAST", "input");
    $alignments = 0;
    while ($alignments !~ /^(y|yes|n|no)$/i) {
        print "\nDoes the BLAST results file contain alignments (y/n) ?:\n";
        chomp ($alignments = <>);
    }
    return 0 if ! &open_file ("filleted BLAST results","OUTFILE", "output");
    print "\n\nFilleting BLAST results ...\n";
    if ($alignments =~/^y/i) {
        &blast_alignments
    }
    else {
        &blast_no_alignments
    }
    close(BLAST);
    close(OUTFILE);
    undef $alignments; undef $query; undef $hit; undef $score; undef $expect; undef @hitline;
    &done(2);

    ####################################

    sub blast_alignments {
        print OUTFILE "Query\tHit\tScore\tExpect(E)\tIdentities\tGaps\tStrand\n";
        while (<BLAST>) {
            chomp;
            if (m/^Query=/) {$query = &get_name($_);}
            elsif (m/^>/) {$hit = &get_name($_);}
            elsif ( m/(?<=^\sScore\s=)\s*\d+(?=\sbits|)/ ) {
                $score = $&;
                $score =~ s/\s//g;
                m/(?<=\sExpect\s=\s).{1,}/;
                $expect = $&;
                $expect =~ s/^e/1e/;
                print OUTFILE "$query\t$hit\t$score\t$expect\t";
                }
            elsif ( m/^\sIdentities\s=\s/ ) {
                if (/Gaps/) {
                    while ( /(\d{1,3}\%)/g ) {
                        print OUTFILE $1."\t";
                    }
                }
                else {
                    $_ =~ /\d{1,3}\%/;
                    print OUTFILE $&."\t\t";
                }
```

```
        }
        elsif (m/(?<=\sStrand\s=\sPlus\s\/\s)(Plus|Minus)/ ) {
            print OUTFILE $&."\n";
        }
    }
}
#####################################

sub blast_no_alignments {
    print OUTFILE "Query\tHit\tScore\tExpect(E)\n";
    while (<BLAST>) {
        chomp;
        if (m/^Query=/) {$query = &get_name($_);}
        elsif (m/Sequences producing significant alignments/) {
            while (<BLAST>) {
                chomp;
                next if /^$/;
                last if /^\s\sDatabase/;
                # s/\s{2,}(?!(\d|e-))/\s/ ;
                # (@hitline) = split(/\s{2,}/,$_);
                /\d{1,}\s{2}[\d\.e-]{1,}$/;
                (@hitline) = split(/\s{2,}/,$&);
                $hit = &get_name($_);
                # $hit = &get_name($hitline[0]);
                print OUTFILE $query."\t".$hit."\t".$hitline[0]."\t".$hitline[1]."\n";
            }
        }
    }
}
}

##############################################################################

sub combine_blast_results {
    my $header = ""; my %blast_files; my $file = 0; my $msg = ""; my $dir_choice = 0;
    while ($dir_choice !~ /^[NCB]$/i) {
        $cons->Cls;
        print "\nOption $option: $options{$option}\n";
        print "\nSelect directory containing BLAST results files:\n";
        print "  (B) - BLAST:    $working_dirs{B}\n";
        print "  (C) - ClustalW:  $working_dirs{C}\n";
        print "  (N) - NCBI:     $working_dirs{N}\n\n";
        chomp ($dir_choice = <>);
        return 0 if $dir_choice=~/^$/;
        $dir_choice = uc $dir_choice;
    }
    chdir ($working_dirs{$dir_choice});
    while ($file !~ /^$/){
        $cons->Cls;
        print "In directory ".cwd."\n";
        print "\nOption $option: $options{$option}\n";
        print $msg;
        print "\nEnter name of BLAST results file " . (scalar (keys %blast_files) + 1);
        print " (just hit return when finished):\n";
        chomp ($file = <>);
        last if $file =~ /^$/;
        $file = lc $file;
        if (! -e $file){$msg = "\nNo file \'$file\' in directory " .cwd. "\nPlease try again\n";}
        elsif (exists ($blast_files{$file})) {
            $msg = "\nYou\'ve already entered file $file. Please try again.\n";
        }
```

```
        else {$msg = "\nFile \'$file\' OK\n"; $blast_files{$file}=1;}
    }
    return 0 if scalar (keys %blast_files) == 0;
    if (scalar (keys %blast_files) == 1) {print"You can\'t combine just one file!\n";sleep 1;return 0};
    return 0 if ! &open_file ("combined BLAST results files","OUTFILE", "output");
    print "\nCombining files...\n";
    for $file (keys %blast_files) {
        return 0 if ! &open_file ("","INFILE","input",$file,cwd);
        while (<INFILE>) {
            if (/^Query\tHit\tScore\t/){
                next if tell OUTFILE != 0;
            }
            print OUTFILE;
        }
        close (INFILE);
    }
    close (OUTFILE);
    &done(2);
}

##########################################################################

sub blast_groups {
    undef %accs; undef %groups; undef $group; undef $acc; undef $groups; undef $max_group;
    undef $self_hits; undef $significance; undef $sigpower; undef $rejected;
    print "\nOption $option: $options{$option}\n";
    print "Reads a file of BLAST results compiled using option H (& possibly I)\n";
    print "and outputs non-redundant groups of homologues\n\n";
    return 0 if ! &open_file ("filleted BLAST results file","BLAST", "input");
    return 0 if ! &open_file ("groups of homologues","OUTFILE", "output");
    $self_hits = 0;
    while ($no_self_hits !~ /^(y|yes|n|no)$/i){
        $cons->Cls;
        print "\nOption $option: $options{$option}\n";
        print "\nIgnore sequences with no homologues other than themselves ? (y/n)\n";
        chomp ($no_self_hits = <>);
    }
    while ($sigpower !~ /^\d+$/i){
        $cons->Cls;
        print "\nOption $option: $options{$option}\n";
        print "\nEnter significance level for BLAST hits:\n1.00E-";
        chomp ($sigpower = <>);
    }

    $significance = 10 ** (-$sigpower);
    while (! eof (BLAST)) {
        &read_blast;
        &combine_groups;
        last if eof (BLAST);
        &tidy_group_data;
    }
    close(BLAST);
    &min_groupsize;
    &output_groups;
    close(OUTFILE);
    undef %accs; undef %groups; undef $group; undef $acc; undef $groups; undef $max_group;
    undef $self_hits; undef $significance; undef $sigpower; undef $rejected;
    &done(2);

    ###################################

    sub read_blast {
```

```
    my $query; my $hit; my $expect; my $lastquery = "";
    print "\nReading input...\n";
    $group = $max_group;
    $. = 0;
    while (<BLAST>) {
        next if /^Query/;
        chomp;
        ($query,$hit,undef,$expect)=split(/\t/,$_);
        next if $no_self_hits =~ /^y/i && $hit eq $query;
        if ($expect=~/^e-/){
            $expect=~s/^e-//;
            if( $sigpower > $expect){
                $rejected++;
                next ;
            }
        }
        elsif ($expect > $significance){
            $rejected++;
            next ;
        }
        if ($lastquery ne $query) {
            $group++;
            $accs{$query}{$group} = 1;
            $groups{$group}{$query} = 1;
        }
        $accs{$hit}{$group} = 1;
        $groups{$group}{$hit} = 1;
        $lastquery = $query;
        $max_group = $group;
        return 0 if ($. == 700000 && ! eof(BLAST) );
    }
}

#####################################

sub combine_groups {
    my $gp_a; my $gp_a_acc; my $gp_b; my $gp_b_acc; my $old_gp_a_size;
    $groups = scalar (keys %groups);
    print "\nCombining groups...\n\n";
GROUP_A:   for $gp_a (keys %groups) {
        $old_gp_a_size = 0;
        while ($old_gp_a_size  < scalar (keys %{$groups{$gp_a}})) {
            if ($old_gp_a_size == 0) {
                print "\n$groups groups left";
            }
            $old_gp_a_size = scalar (keys %{$groups{$gp_a}});
            for $gp_a_acc (keys %{$groups{$gp_a}}) {
GROUP_B:                for $gp_b (keys %{$accs{$gp_a_acc}}) {
                    next GROUP_B if $gp_a eq $gp_b;
                    next GROUP_B if ! defined (%{$groups{$gp_b}});
                    for $gp_b_acc (keys %{$groups{$gp_b}}) {
                        $groups{$gp_a}{$gp_b_acc} = 1;
                    }
                    undef (%{$groups{$gp_b}});
                    $groups --;
                }
                undef (%{$accs{$gp_a_acc}});
            }
        }
    }
}
```

```
    ###################################
    sub tidy_group_data {
        print "\nTidying data\n";
        undef %accs;
        my %temp_groups;
        for $group (keys %groups) {
            next if scalar (keys %{$groups{$group}}) == 0;
            %{$temp_groups{$group}} = %{$groups{$group}};
            for $acc (keys %{$groups{$group}}) {
                $accs{$acc}{$group}=1;
            }
        }
        undef %groups;
        %groups = %temp_groups;
    }
    ###################################
    sub min_groupsize {
        my $happy = "n"; my ($temp_groups, $min_gp_size);
        while ( $happy !~ /^(y|yes)$/i ) {
            $cons->Cls;
            if ($no_self_hits =~ /^n/i){print "\n$groups groups of one or more homologues found\n";}
            else {print "\n$groups groups of two or more homologues found\n";}
            print "\nWhat is the minimum number of homologues required in a group?\n";
            chomp ($min_gp_size = <> );
            $temp_groups = 0;
            for $group (keys %groups) {
                if(scalar(keys %{$groups{$group}})>0 && scalar(keys %{$groups{$group}})>=
$min_gp_size) {
                    $temp_groups++;
                }
            }
            print "A minimum of $min_gp_size homologues gives $temp_groups groups\n";
            print "Is this OK? (y/n)";
            chomp ($happy = <> );
        }
        for $group (keys %groups) {
            if (scalar (keys %{$groups{$group}}) < $min_gp_size) {undef %{$groups{$group}}; }
        }
    }
    ###################################
    sub output_groups {
        my $i=0;
        print "\n\nOutputting groups...\n";
        for $group (keys %groups) {
            next if ! defined %{$groups{$group}};
            $i++;
            print OUTFILE "Group$i\n";
            for $acc (keys %{$groups{$group}}) {
                print OUTFILE $acc . "\n";
            }
            print OUTFILE "\n";
        }
    }
}
#############################################################################
```

```
sub extract_groups {
    undef %long_seqs; undef $long_seq; undef %groups; undef %accessions;
    undef %seq_lengths; undef $group; undef $seq;
    print "\nOption $option: $options{$option}\n";
    return 0 if ! &open_file ("list of groups of homologous sequences","GROUPS", "input");
    &read_groups (\%accessions, \%groups);
    close(GROUPS);
    return 0 if ! &open_file ("database of sequences","SEQFILE", "input");
    return 0 if ! &open_file ("longest sequences in each group","LONGSEQS", "output");
    return 0 if ! &open_file ("groups to make new BLAST database","GROUPS_DB", "output");
    &get_seq_lengths;
    &find_long_seqs;
    seek SEQFILE, 0, 0;
    &output_seqs;
    close(SEQFILE);
    close(LONGSEQS);
    close (GROUPS_DB);
    undef %long_seqs; undef $long_seq; undef %groups; undef %accessions;
    undef %seq_lengths; undef $group; undef $seq;
    &done(2);

    ####################################

    sub get_seq_lengths {
        my $len; $seq = "";
        print "\nFinding sequence lengths...\n";
        while (<SEQFILE>) {
            if(/^>/) {
                if ($seq ne "") {$seq_lengths{$seq}=$len;}
                $seq = &get_name($_);
                $len = 0;
            }
            else {
                $len += length $_;
            }
        }
        $seq_lengths{$seq}=$len;
    }

    ####################################

    sub find_long_seqs {
        my ($long_seq_len);
        print "\nFinding longest sequences in groups...\n";
        for $group (keys %groups) {
            $long_seq_len = 0;
            for $seq (keys %{$groups{$group}}) {
                if ($seq_lengths{$seq} > $long_seq_len) {
                    $long_seq_len = $seq_lengths{$seq};
                    $long_seq = $seq;
                }
            }
            $long_seqs{$long_seq} = 1;
        }
    }

    ####################################

    sub output_seqs {
        my $name; my $output_seq = 0;
        print "\nOutputting sequences...\n";
        while (<SEQFILE>) {
```

```
        if(m/^>/) {
            chomp;
            $name = &get_name($_);
            if (exists ($long_seqs{$name})) {
                delete $long_seqs{$name};
                delete $accessions{$name};
                $long_seq = 1;
            }
            elsif (exists ($accessions{$name})) {
                $output_seq = 1;
                $long_seq = 0;
                delete $accessions{$name};
            }
            else {$long_seq = 0; $output_seq = 0;}
        }
        if ($long_seq) {print GROUPS_DB; print LONGSEQS;}
        elsif ($output_seq) {print GROUPS_DB;}
    }
  }
}
##########################################################################
sub read_groups {
    my ($acc_hash_ref, $group_hash_ref) = @_; my $group = 0;
    print "Reading groups...\n";
    while (<GROUPS>) {
        chomp;
        if (m/^Group\d+$/) {
            $group++;
            next;
        }
        next if /^$/;
        ${$group_hash_ref}{$group}{$_} = 1;
        ${$acc_hash_ref}{$_} = $group;
    }
}
##########################################################################

sub create_seqfiles {
    chdir $clustal_dir;
    undef %acc_gp; undef %groups; undef %long_seqs; undef $group; undef $groups;
    undef $i; undef $outfile; undef %outfiles; undef $seq; undef $acc;
    my @old_files; my $old_file;
    opendir (DIR, "$clustal_dir") || die " can not read directory $clustal_dir";
    @old_files = sort grep(/^Gp\d{5}\.seq$/i, readdir(DIR));
    closedir DIR;
    if ($#old_files>=0) {
        print "\nThere are ". ($#old_files +1) . ' old Gp*.seq files in the ClustalW directory.';
        print "\nThese must be transferred to another directory or deleted.\n";
        print "\nDo you want to delete the files? (y/n) \n";
        chomp ($response = <>);
        if ($response =~ /^(y|yes)$/i) {
            chdir $clustal_dir;
            for $old_file (@old_files) {
                unlink $old_file;
            }
            print "Files deleted\n";
        }
        else {print "\nPlease move the files before proceeding."; sleep 2; return 0}
    }
```

```
return 0 if ! &open_file ("list of groups of homologous sequences","GROUPS", "input");
&read_groups (\%acc_gp, \%groups);
close(GROUPS);
return 0 if ! &open_file ("long sequence BLAST results file","BLAST", "input");
print "Reading BLAST file...\n";
&read_blast_strands;
close(BLAST);
return 0 if ! &find_missing_hits;
return 0 if ! &open_file ("database of homologous sequence groups ","SEQFILE", "input");
$groups = scalar (keys %groups);
&min_groupsize;
&max_groupsize;
chdir $clustal_dir;
&output_clustal_files;
close(SEQFILE);
$cons->Cls;
print "Creating list of Gp*.seq files in ClustalW directory for alignment\n\n";
print "The list is required for launching ClustalW for multiple alignments - option N\n";
print "The list can also be created using option M\n";
return 0 if ! &open_file("","OUTFILE","output","Seqfiles.txt",$clustal_dir);
for $outfile (sort keys %outfiles) {print OUTFILE "$outfile\n";}
close (OUTFILE);
undef %acc_gp; undef %groups; undef %long_seqs; undef $group; undef $groups;
undef $i; undef $outfile; undef %outfiles; undef $seq; undef $acc;
&done(4);

####################################

sub read_blast_strands {
    my ($query,$hit,$strand);
    $_=<BLAST>;
    while (<BLAST>) {
        chomp;
        ($query,$hit,undef,undef,undef,undef,$strand)=split(/\t/,$_);
        $long_seqs{$query} = 1; #To be retained in .seq file if group size is to be reduced
        $group = $acc_gp{$query};
        $strand = "Plus" if $hit eq $query;
        if (exists($groups{$group}{$hit})) {
            $groups{$group}{$hit} = $strand;
        }
    }
}

####################################

sub find_missing_hits {
    my $missing_hits = 0;
    chdir $blast_dir;
    return 0 if ! &open_file("","MISSING","output","Missing_Hits.txt",$blast_dir);
    print MISSING "Group\tGP size\tAccession\tStrand\n";
    for $group (sort {$a<=>$b} keys %groups) {
        for $acc (keys %{$groups{$group}}) {
            if ($groups{$group}{$acc} !~ /(Minus|Plus)/i) {
                print MISSING "$group\t". scalar (keys %{$groups{$group}});
                print MISSING "\t$acc\t$groups{$group}{$acc}\n";
                $missing_hits++;
            }
        }
    }
    close (MISSING);
    if ($missing_hits){
        $cons->Cls;
```

```
        print "\nStrand orientations missing for $missing_hits sequences\n";
        print "(See file \'Missing_Hits.txt\' in $blast_dir for details)\n";
        print "\nThis may cause sequences to be outputted in the wrong orientation for\n";
        print "alignment by Clustal\n";
        print "\nRepeating the BLAST with a lower E value (option -e) and returning a \n";
        print "higher number of hits & alignments per sequence (options -b and -v)\n";
        print "may resolve the problem\n";
        $response = 0;
        while ($response !~ /^(y|yes|n|no|x|exit)$/){
            print "\nRemove sequences for which strand orientation is unknown? (y/n/x -exit)\n";
            chomp ($response = <>);
        }
        return 0 if $response =~/^(x|exit)$/;
        if ($response =~/^y/){
            for $group (keys %groups){
                for $acc (keys %{$groups{$group}}){
                    if ($groups{$group}{$acc} !~ /^(Minus|Plus)$/){
                        delete $groups{$group}{$acc};
                        delete $acc_gp{$acc};
                    }
                }
            }
        }
    }
    return 1;
}

####################################

sub max_groupsize {
    my (@temp, $max_gp_size, $kill_seq);
    print "\nWhat is the maximum number of homologues required in a group?\n";
    chomp ($max_gp_size = <>);
    for $group (keys %groups) {
        next if ! defined %{$groups{$group}};
        foreach $acc (keys %{$groups{$group}}) {
            if (! exists ($long_seqs{$acc})) {
                push @temp, $acc;
            }
        }
        while (scalar (keys %{$groups{$group}}) > $max_gp_size) {
            $kill_seq = (rand, ($#temp - 0.001));
            $kill_seq = int $kill_seq;
            delete $groups{$group}{$temp[$kill_seq]};
            delete $acc_gp{$temp[$kill_seq]};
            splice (@temp, $kill_seq, 1);
        }
        $#temp = -1;
    }
}

####################################

sub output_clustal_files {
    print "\nOutputting groups of sequences...\n";
    OUTER: while (<SEQFILE>) {
        if(m/^>/) {
            $acc = &get_name($_);
            next OUTER if ! exists ($acc_gp{$acc});
            $group = $acc_gp{$acc};
            next OUTER if ! defined ( %{$groups{$group}} );
            if (exists($groups{$group}{$acc})) {
```

```
                $outfile = "Gp" . ("0" x (5-(length $group))) . $group . ".seq";
                $outfiles{$outfile}=1;
                return 0 if ! &open_file("",OUTFILE,"append",$outfile,$clustal_dir);
                print OUTFILE "\>$acc\n";
                while (<SEQFILE>) {
                    last if m/^>/;
                    chomp ($seq .= $_);
                }
                if ($groups{$group}{$acc} eq "Minus") {&rev_comp;}
                while (length $seq > 0) {print OUTFILE substr($seq,0,70,"") . "\n";}
                close (OUTFILE);
                redo OUTER;
            }
            else {
                next OUTER;
            }
        }
    }

    ###################################

    sub rev_comp{
        my $revseq = "";
        for ($i=length($seq); $i >= 0; $i--) {
            $revseq .= substr($seq,$i,1);
        }
        $seq=$revseq;
        $seq =~
tr/tcagnrymkswvdbhTCAGNRYMKSWVDBH/agtcnyrkmwsbhvdAGTCNYRKMWSBHVD/;
    }
}

#############################################################################

sub clustal_filelist {
    undef @files; undef $file;
    print "\nOption $option: $options{$option}\n";
    chdir $clustal_dir;
    &get_filenames;
    print "There are " . ($#files + 1) . " files to align\n";
    return 0 if ! &open_file("","OUTFILE","output","Seqfiles.txt",$clustal_dir);
    foreach $file (@files) {print OUTFILE "$file\n";}
    close (OUTFILE);
    undef @files; undef $file;
    &done(3);

    ###################################

    sub get_filenames{
        opendir (DIR, "$clustal_dir") || die " unable to read directory $clustal_dir";
        @files = sort grep(/^Gp\d{5}\.seq$/i, readdir(DIR));
        closedir DIR;
    }
}

#############################################################################

sub batch_clustal {
    use File::Copy; use POSIX qw(ceil);
    undef @filelist; undef @files; undef $file; undef $aln_dir;
    my $batch_no =0; my $max_batch_no =0;
```

```
    $aln_dir = $clustal_dir . '\SNPalignments';
    print "Alignment files will be outputted to $aln_dir\n";
    print "\nOption $option: $options{$option}\n";
    if (!-e $aln_dir || ! -d $aln_dir) {mkdir $aln_dir, 0744;}
    chdir $clustal_dir;
    if (!-e "ClustalW.exe") {
        print 'You must have ClustalW.exe in the ClustalW working directory!'."\n";
        print "Either put this program into directory $clustal_dir\n";
        print "or change the working directory to one that contains ClustalW.";
        sleep 4; return 0;
    }
    if (! -e "Seqfiles.txt" ) {
        print "\nYou need a file called \'SeqFiles.txt\' listing the .seq files in the directory\n";
        print "\'$clustal_dir\'. This file must also be in $clustal_dir.\n";
        print "To create this file, run option M from the menu.\n";
        sleep 4;return 0;
        }
    return 0 if ! &open_file("","FILELIST","input","Seqfiles.txt",$clustal_dir);
    while (<FILELIST>) {
        chomp;
        push @filelist, $_;
        }
    close (FILELIST);
    $max_batch_no = POSIX::ceil($#filelist/50);

    while ($batch_no =~/\D/ || $batch_no < 1 || $batch_no > $max_batch_no ) {
        $cons->Cls;
        print "There are " . ($#filelist +1) ." files to align\n";
        return 0 if $#filelist < 0;
        print "\nEnter the number of the batch of 50 sequence files to align:\n";
        print "(Between 1 and $max_batch_no)\n";
        chomp ($batch_no = <>);
        return 0 if $batch_no =~ /^$/;
    }
    @files = splice (@filelist, ($batch_no-1)*50,50);
    print "\n";
    &align;
    undef @filelist; undef @files; undef $file; undef $aln_dir;
    &done(2);

    ####################################

    sub align{
        my $no_files = $#files+1;
        foreach $file (@files) {
            print "Aligning $file\tFiles left: $no_files\n";
            $no_files --;
            `clustalw /infile=$file /quicktree`;
            $file =~ s/.seq//;
            if (move("$file.aln","$aln_dir\\$file.aln")) {
                move("$file.seq","$aln_dir\\$file.seq");
                }
            unlink "$file.dnd";
        }
    }
}

##########################################################################

sub get_snps {
    undef @files; undef $response; undef $aln_dir; undef $cutoff; undef %counts;
    undef $snp_threshold;
```

```
$aln_dir = $clustal_dir.'\SNPalignments';
chdir $aln_dir;
$response = -1;
while ($response !~ /^(y|n|yes|no)$/i ) {
    $cons->Cls;
    print "\nOption $option: $options{$option}\n";
    print "\nAre alignment files in the directory\n$aln_dir ? (y/n)\n";
    chomp ($response  = <> );
    return 0 if $response =~/^$/;
}
if ($response=~/^n/i) {
    $aln_dir =-1;
    while (! -e $aln_dir) {
        print "\nEnter the name of the directory containing alignment files:\n";
        chomp  ($aln_dir = <>);
        return 0 if $aln_dir =~ /^$/;
    }
}
&read_dir ($aln_dir);
print "$aln_dir\n";
foreach $file (@files) {print "$file\n";}
print "\n" . ($#files + 1) . " alignment files\n";
sleep 1;
return 0 if ! &open_file ("SNP results","SNPFILE", "output");
print SNPFILE "Alignment\tBase\tG\tA\tT\tC\tN\t-\tOther\n\n";
return 0 if ! &open_file("","RESULTSTAB", "output","SNPresults.out",cwd);
print RESULTSTAB "Alignment\tSNPs\n";
return 0 if ! &open_file ("consensus sequences","CONSFILE", "output");
$snp_threshold = 0;
while ($snp_threshold!~/^(\d+|0?\.\d+)$/ || $snp_threshold <= 0){
    $cons->Cls;
    print "\nOption $option: $options{$option}\n";
    print "\nWhat level of the minor base is required to define an SNP?\n";
    print "\tFor a minimum frequency, type a fraction from 0 to 0.5\n";
    print "\tFor a minimum absolute number, type an integer.\n";
    chomp ($snp_threshold = <>);
    return 0 if $snp_threshold =~/^$/;
    if ($snp_threshold > 0.5 && $snp_threshold < 1){$snp_threshold = 0;}
}
if ($snp_threshold < 1){
    $ignore_lone_snps = 0;
    while ($ignore_lone_snps !~ /^(y|yes|n|no)$/i){
        $cons->Cls;
        print "\nOption $option: $options{$option}\n";
        print "\nSNP threshold frequency: $snp_threshold\n";
        print "\nDo you want to consider SNPs present as a single copy as artefacts? (y/n)\n";
        chomp ($ignore_lone_snps = <>);
        return 0 if $ignore_lone_snps =~/^$/;
    }
}
$cutoff = "";
while ($cutoff !~ /^\d+$/){
    $cons->Cls;
    print "\nOption $option: $options{$option}\n";
    print "\nHow many bases at the start and end of sequences should be ignored?\n";
    chomp ($cutoff = <> );
    return 0 if $cutoff =~/^$/;
}
if ($cutoff==0) {$cutoff=-1;}
foreach $aln_file (@files) {
    chdir "$aln_dir";
    return 0 if ! &open_file("","ALIGNMENT","input",$aln_file,$aln_dir);
```

```
    $group = $aln_file;
    $group =~ s/.aln//;
    &snp_list;
    if ($snps>0) {print SNPFILE "\n";}
    close (ALIGNMENT);
}
close (SNPFILE);
close (CONSFILE);
close (RESULTSTAB);
undef @files; undef $response; undef $aln_dir; undef $cutoff; undef %counts;
undef $snp_threshold;
&done(2);

####################################

sub read_dir{
    my $dir = shift @_;
    opendir (DIR, "$dir") || die " can not read directory $dir";
    @files = sort grep(/\.aln$/, readdir(DIR));
    closedir DIR;
}

####################################

sub snp_list {
    $#aln = -1;
    $_ = <ALIGNMENT>;
    $i = 0;
    while (<ALIGNMENT>) {
        if (/^\s*$/ || /^\s+\*/) {$i = 0;}
        else {
            s/^\S+\s+//;
            chomp;
            $aln[$i] .= $_;
            $i++;
        }
    }
    for ($i= 0; $i<=$#aln; $i++) {
        $base = 0;
START:      for ($j=0; $j<(length $aln[$i]);$j++) {
            if (substr ($aln[$i], $j, 1) ne '-') {
                $base++;
                if ($base >= $cutoff) {
                    $start[$i] = $j;
                    last START;
                }
            }
        }
        $base = 0;
END:    for ($j=(length $aln[$i])-1; $j>0;$j--) {
            if (substr ($aln[$i], $j, 1) ne '-') {
                $base++;
                if ($base >= $cutoff) {
                    $end[$i] = $j;
                    last END;
                }
            }
        }
    }
    foreach $seq (@aln) {
        for ($i=0; $i<(length $seq);$i++) {
            last if substr ($seq, $i, 1) ne '-';
```

```
            substr ($seq, $i, 1, "z");
        }
        for ($i=(length $seq)-1; $i>0;$i--) {
            last if substr ($seq, $i, 1) ne '-';
            substr ($seq, $i, 1, "z");
        }
    }
    $seqs = $#aln +1;
    if ($snp_threshold < 1) {$snp_def = $seqs * $snp_threshold;}
    else {$snp_def = $snp_threshold;}
    if ($ignore_lone_snps =~ /^y/i){
        if ($snp_def < 2) {$snp_def = 2;}
    }
    $snps = 0;
    $consens_seq = "";
    for ($i=0; $i < length $aln[0]; $i++) {

$counts{g}=0;$counts{a}=0;$counts{t}=0;$counts{c}=0;$counts{n}=0;$counts{o}=0;$counts{x}=0;
SEQ:    for ($j=0; $j < $seqs; $j++) {
            next SEQ if ($i < $start[$j] || $i > $end[$j]);
            $base = lc (substr ($aln[$j],$i,1));
            $base =~ tr/gatcnz\-^[gatcnz\-]/gatcnzox/;
            $counts{$base}++;
        }
        $bases = 0;
        if ($counts{g} >= $snp_def ) {$bases++;}
        if ($counts{a} >= $snp_def ) {$bases++;}
        if ($counts{t} >= $snp_def ) {$bases++;}
        if ($counts{c} >= $snp_def ) {$bases++;}
        if ($bases>=2) {
            print SNPFILE "$group\t".($i+1)."\t$counts{g}\t$counts{a}\t$counts{t}\t$counts{c}\t";
            print SNPFILE "$counts{n}\t$counts{o}\t$counts{x}\n";
            $snps++;
        }
        $consens_seq .= &consensus_base;
    }
    print RESULTSTAB "$aln_file\t$snps\n";
    if ($snps == 1){print "$aln_file\t1 SNP\n";}
    else {print "$aln_file\t$snps SNPs\n";}
    print CONSFILE '>'."$group\n";
    while (length $consens_seq > 0) {
        print CONSFILE substr($consens_seq,0,70,"") . "\n";
    }
}
#####################################

sub consensus_base {
    my ($consens_base, $ntide);
    my $max_count = 0;
    foreach $ntide (keys %counts) {
        if ($ntide =~ /[gatcno]/) {
            if ($max_count < $counts{$ntide}) {
                $max_count = $counts{$ntide};
                $consens_base = $ntide;
            }
            elsif ($max_count == $counts{$ntide}) {
                $consens_base = "n";
            }
        }
    }
    if ($consens_base eq "o") {$consens_base = "";}
    return $consens_base;
```

```
    }
}

#########################################################################

sub get_name {
    my $search_str = shift @_;
    if ( $search_str =~ /\|/ ) {
        for $db_id (@db_ids) {
            if ( m/(?<=$db_id\|)[^\|]{1,}(?=\|)/ ) {return $&;}
            elsif ( m/(?<=$db_id\|\|)[^\s]{1,}(?=\s)/ ) {return $&;}
            elsif ( m/(?<=$db_id\|).{1,20}(?=\s)/ ) {return $&;}
        }
    }
    else {
        $search_str =~ s/^>//;
        $search_str =~ s/^Query=\s//;
        return $search_str;
    }
}

#########################################################################

sub prog_setup {
    chdir $working_dirs{P};
    my $setup_option = 0;
    my ($new_clustal_dir, $new_blast_dir, $new_ncbi_dir);
    my (@new_db_ids, $db_ids, $new_db_ids);
    $db_ids = join (' ', @db_ids);
    $db_ids =~ s/\s$//;
OUTER:  while() {
        $setup_option = 0;
        while ($setup_option !~ /^[1-4mMxX]$/) {
            &setup_menu;
            chomp ($setup_option = <>);
        }
        exit 0 if $setup_option =~ /^x$/i;
        return if $setup_option =~ /^m$/i;
        if ($setup_option == 1) {
            @new_db_ids = @db_ids;
            $input = -1;
    DBIDMENU:    while ($input !~ /^$/ ) {
                $cons->Cls;
                print "\Database identifiers currently recognised:\n";
                for ($i=0;$i<=$#new_db_ids;$i++) {
                    if ($i % 5 == 0) {print "\n";}
                    print $i+1 . ": $new_db_ids[$i]\t";
                }
                print "\n\nEnter database identifier number to remove or type new database identifier (Hit
Return to continue):\n";
                chomp ($input = <>);
                last DBIDMENU if $input =~ /^$/;
                if ($input =~ /^\d+$/) {
                    splice (@new_db_ids,$input-1,1);
                    next DBIDMENU;
                    }
                else {
                    for $db_id (@new_db_ids) {
                        if ($db_id eq $input) {
                            print "\nDatabase identifier $input already exists!\n";<>;
                            next DBIDMENU;
                        }
```

```
            }
            push (@new_db_ids,$input);
        }
        }
        $new_db_ids = join (' ', @new_db_ids);
        $new_db_ids =~ s/\s$//;
        next OUTER if $new_db_ids eq $db_ids;
        $db_ids = $new_db_ids;
        @db_ids = split (/\s/, $db_ids);
    }
    elsif ($setup_option == 2) {
        $new_blast_dir = &new_dir ("\nBLAST working directory:",$blast_dir);
        next OUTER if ! $new_blast_dir;
        $blast_dir = $new_blast_dir;
    }
    elsif ($setup_option == 3) {
        $new_clustal_dir = &new_dir ("\nClustalW working directory:",$clustal_dir);
        next OUTER if ! $new_clustal_dir;
        $clustal_dir = $new_clustal_dir;
    }
    elsif ($setup_option == 4) {
        $new_ncbi_dir = &new_dir ("\nNCBI working directory:",$ncbi_dir);
        next OUTER if ! $new_ncbi_dir;
        $ncbi_dir = $new_ncbi_dir;
    }
    return 0 if ! &open_file("","SETUP","output","SNP_Hunter_Setup.txt",$working_dirs{P});
    print SETUP '@db_ids = split (/\s/,' . "\'$db_ids\')\n";
    print SETUP '$blast_dir = ' . "\'$blast_dir\'\n";
    print SETUP '$clustal_dir = ' . "\'$clustal_dir\'\n";
    print SETUP '$ncbi_dir = ' . "\'$ncbi_dir\'\n";
    close (SETUP);
    print "\nUpdated. Press Return to continue.\n";<>;
}

####################################

sub setup_menu {
    $cons->Cls;
    print "\n\tSNP_HUNTER SETTINGS MENU\n";
    print "\n  Please select an option:\n\n";
    print "  1\tUpdate recognised sequence database identifiers\n";
    print "  2\tChange working directory for BLAST\n";
    print "  3\tChange working directory for ClustalW\n";
    print "  4\tChange working directory for NCBI downloads\n";
    print "\n";
    print "  M\tReturn to main menu\n";
    print "  X\tExit program\n\n";
    print "  Option: ";
}

####################################

sub new_dir{
    my ($prompt, $old_dir) = @_;
    my $new_dir=""; my $err = "";
    while (! -e $new_dir || ! -d $new_dir) {
        $cons->Cls;
        print $err;
        print "$prompt $old_dir\n\n";
        print "Enter new directory:\n";
        chomp ($new_dir = <>);
        return 0 if $new_dir =~ /^$/;
```

```
            $err = "\nDirectory $new_dir does not exist\nPlease try again\n";
        }
        return 0 if $new_dir eq $old_dir;
        return $new_dir;
    }
}

#########################################################################

sub open_file {
    my ($prompt, $filehandle, $file_type, $filename, $file_dir) = @_;
    $filename ||= 0; $file_dir ||= 0;
    my $err = ""; my $dir_choice = 0; my $dir_txt = "";
    if ($prompt!~/^$/){
        if ($file_type eq "output") {$prompt = "\nEnter filename for outputting $prompt:\n";}
        else {$prompt = "\nEnter the filename of the $prompt:\n";}
        while (! -e $filename || ! -f $filename) {
            $cons->Cls;
            print "\nOption $option: $options{$option}\n";
            print $err;
            print $prompt;
            chomp ($filename = <> );
            return 0 if $filename =~ /^$/;
            $dir_choice=0;
            while ($dir_choice !~ /^[NCB]$/i){
                $cons->Cls;
                print "\nOption $option: $options{$option}\n";
                if ($file_type ne "output"){print "\nWhich directory is file $filename in ?\n";}
                else {print"\nWhich directory do you want to output file $filename to ?\n";}
                print "  (B) - BLAST:    $working_dirs{B}\n";
                print "  (C) - ClustalW: $working_dirs{C}\n";
                print "  (N) - NCBI:     $working_dirs{N}\n\n";
                chomp ($dir_choice = <>);
                return 0 if $dir_choice=~/^$/;
            }
            $file_dir = $working_dirs{uc($dir_choice)};
            last if $file_type eq "output";
            chdir $file_dir;
            if ($file_type =~ /(input|append)/){
                $dir_txt = $file_dir;
                $dir_txt =~ s/\/\\/g;
                if (! -e $filename) {
                    $err = "\nNo file \'$filename\' in directory $dir_txt\nPlease try again\n";
                }
            }
        }
    }
    chdir $file_dir;
    $current_dir = cwd;
    if ($file_type eq "output") {open($filehandle,">$filename") || die "opening $filename: $!"; $_="1";}
    elsif ($file_type eq "input") {open ($filehandle, "<$filename") || die "opening $filename: $!"; $_="1";}
    elsif ($file_type eq "append") {open ($filehandle, ">>$filename") || die "opening $filename: $!";
$_="1";}
    $open_filehandles{$filehandle}=1;
    return 1;
}

#########################################################################
```

LITERATURE CITED

LITERATURE CITED

Agarwala, R., D. L. Applegate, D. Maglott, G. D. Schuler and A. A. Schaffer (2000)
A fast and scalable radiation hybrid map construction and integration strategy. Genome Res **10:** 350-64.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang *et al.* (1997)
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-402.

Amarante, M. R., Y. P. Yang, S. R. Kata, C. R. Lopes and J. E. Womack (2000)
RH maps of bovine chromosomes 15 and 29: conservation of human chromosomes 11 and 5. Mamm Genome **11:** 364-8.

Amills, M., V. Ramiya, J. Norimine and H. A. Lewin (1998)
The major histocompatibility complex of ruminants. Rev Sci Tech **17:** 108-20.

Andersson, L., A. Lunden, S. Sigurdardottir, C. J. Davies and L. Rask (1988)
Linkage relationships in the bovine MHC region. High recombination frequency between class II subregions. Immunogenetics **27:** 273-80.

Ashworth, L. K., M. A. Batzer, B. Brandriff, E. Branscomb, P. de Jong *et al.* (1995)
An integrated metric physical map of human chromosome 19. Nat Genet **11:** 422-7.

Bachtrog, D., M. Agis, M. Imhof and C. Schlotterer (2000)
Microsatellite variability differs between dinucleotide repeat motifs-evidence from Drosophila melanogaster. Mol Biol Evol **17:** 1277-85.

Band, M., J. H. Larson, J. E. Womack and H. A. Lewin (1998)
A radiation hybrid map of BTA23: identification of a chromosomal rearrangement leading to separation of the cattle MHC class II subregions. Genomics **53:** 269-75.

Band, M. R., J. H. Larson, M. Rebeiz, C. A. Green, D. W. Heyen *et al.* (2000)
An ordered comparative map of the cattle and human genomes. Genome Res **10:** 1359-68.

Barendse, W., S. M. Armitage, L. M. Kossarek, A. Shalom, B. W. Kirkpatrick *et al.* (1994)
A genetic linkage map of the bovine genome. Nat Genet **6:** 227-35.

Barendse, W., D. Vaiman, S. J. Kemp, Y. Sugimoto, S. M. Armitage *et al.* (1997)
A medium-density genetic linkage map of the bovine genome [published erratum appears in Mamm Genome 1997 Oct;8(10):798]. Mamm Genome **8:** 21-8.

Barrett, J. H. (1992)
Genetic mapping based on radiation hybrid data. Genomics **13:** 95-103.

Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste and F. Bonhomme (2001)
Genetix, logiciel sous Windows TM pour la génétique des populations, Université de Montpellier II, Montpellier (France).

Bergstrom, T. F., H. Engkvist, R. Erlandsson, A. Josefsson, S. J. Mack *et al.* (1999)
Tracing the Origin of HLA-DRB1 Alleles by Microsatellite Polymorphism. Am J Hum Genet **64:** 1709-1718.

Bertorelle, G., and L. Excoffier (1998)
Inferring admixture proportions from molecular data. Mol Biol Evol **15:** 1298-311.

Bishop, M. D., S. M. Kappes, J. W. Keele, R. T. Stone, S. L. Sunden *et al.* (1994)
A genetic linkage map for cattle. Genetics **136:** 619-39.

Boehnke, M., K. Lange and D. R. Cox (1991)
Statistical methods for multipoint radiation hybrid mapping. Am J Hum Genet **49:** 1174-88.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd *et al.* (1994)
High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455-7.

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:** 5131-5.

Brezinsky, L., S. J. Kemp and A. J. Teale (1992)
ILSTS001: a polymorphic bovine microsatellite. Anim Genet **23:** 81.

Brezinsky, L., S. J. Kemp and A. J. Teale (1993a)
Five polymorphic bovine microsatellites (ILSTS010-014). Anim Genet **24:** 75-6.

Brezinsky, L., S. J. Kemp and A. J. Teale (1993b)
ILSTS005: a polymorphic bovine microsatellite. Anim Genet **24:** 73.

Brezinsky, L., S. J. Kemp and A. J. Teale (1993c)
ILSTS006: a polymorphic bovine microsatellite. Anim Genet **24:** 73.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne and B. Rolf (1998)
Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet **62:** 1408-15.

Buitkamp, J., F. W. Schwaiger, S. Solinas-Toldo, R. Fries and J. T. Epplen (1995)
The bovine interleukin-4 gene: genomic organization, localization, and evolution. Mamm Genome **6:** 350-6.

Carter, P. L., and J. D. Clark (1976)
Adrar Bous and African cattle, pp. 487-493 in *The 7th Pan-African Congress on Prehistory, 1971*, Addis Ababa.

Chakraborty, R. (1990)
Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. Am J Hum Genet **47:** 87-94.

Chakraborty, R., M. I. Kamboh, M. Nwankwo and R. E. Ferrell (1992)
Caucasian genes in American blacks: new data. Am J Hum Genet **50:** 145-55.

Charlesworth, B., M. T. Morgan and D. Charlesworth (1993)
The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-303.

Charlesworth, B., M. Nordborg and D. Charlesworth (1997)
The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research **70:** 155-174.

Chen, F. C., and W. H. Li (2001)
Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet **68:** 444-56.

Chowdhary, B. P., L. Fronicke, I. Gustavsson and H. Scherthan (1996)
Comparative analysis of the cattle and human genomes: detection of ZOO-FISH and gene mapping-based chromosomal homologies. Mamm Genome **7:** 297-302.

Clark, A. G. (1987)
Neutrality tests of highly polymorphic restriction-fragment-length polymorphisms. Am J Hum Genet **41:** 948-56.

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:** 2001-14.

Cox, D. R., M. Burmeister, E. R. Price, S. Kim and R. M. Myers (1990)
Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science **250:** 245-50.

Crawford, A. M., K. G. Dodds, A. J. Ede, C. A. Pierson, G. W. Montgomery *et al.* (1995)
An autosomal genetic linkage map of the sheep genome. Genetics **140:** 703-24.

Creighton, P., A. Eggen, R. Fries, S. A. Jordan, J. Hetzel *et al.* (1992)
Mapping of bovine markers CYP21, PRL, and BOLA DRBP1 by genetic linkage analysis in reference pedigrees. Genomics **14:** 526-8.

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

Davies, C. J., I. Joosten, L. Andersson, M. A. Arriens, D. Bernoco *et al.* (1994)
Polymorphism of bovine MHC class II genes. Joint report of the Fifth International Bovine Lymphocyte Antigen (BoLA) Workshop, Interlaken, Switzerland, 1 August 1992. Eur J Immunogenet **21:** 259-89.

Deloukas, P., G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund *et al.* (1998)
A physical map of 30,000 human genes. Science **282:** 744-6.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.* (1994)
Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A **91:** 3166-70.

d'Ieteren, G. D., E. Authie, N. Wissocq and M. Murray (1998)
Trypanotolerance, an option for sustainable livestock production in areas at risk from trypanosomosis. Rev Sci Tech **17:** 154-75.

Eggen, A., and R. Fries (1995)
An integrated cytogenetic and meiotic map of the bovine genome. Anim Genet **26:** 215-36.

Eggen, A., S. Solinas-Toldo, A. B. Dietz, J. E. Womack, G. Stranzinger *et al.* (1992)
RASA contains a polymorphic microsatellite and maps to bovine syntenic group U22 on chromosome 7q2.4-qter. Mamm Genome **3:** 559-63.

Ellegren, H., C. J. Davies and L. Andersson (1993)
Strong association between polymorphisms in an intronic microsatellite and in the coding sequence of the BoLA-DRB3 gene: implications for microsatellite stability and PCR-based DRB3 typing. Anim Genet **24:** 269-75.

Epstein, H. (1971)
The Origin of the Domestic Animals of Africa. Africana, New York.

Epstein, H., and I. L. Mason (1984)
Cattle, pp. 6-27 in *Evolution of Domesticated Animals*, edited by I. L. Mason. Longman, London.

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

Felius, M. (1995)
Cattle Breeds - an Encyclopedia. Misset, Doetinchem, Netherlands.

Felsenstein, J. (2001)
PHYLIP (the PHYLogeny Inference Package), University of Washington, USA.

Ferretti, L., P. Leone, F. Pilla, Y. Zhang, M. Nocart *et al.* (1994)
Direct characterization of bovine microsatellites from cosmids: polymorphism and synteny mapping. Anim Genet **25:** 209-14.

Figueroa, F., O. h. C, H. Tichy and J. Klein (1994)
The origin of the primate Mhc-DRB genes and allelic lineages as deduced from the study of prosimians. J Immunol **152:** 4455-65.

Fries, R., A. Eggen and J. E. Womack (1993)
The bovine genome map. Mamm Genome **4:** 405-28.

Fries, R., R. Hediger and G. Stranzinger (1988)
The loci for parathyroid hormone and beta-globin are closely linked and map to chromosome 15 in cattle. Genomics **3:** 302-7.

Frisch, J. E., R. Drinkwater, B. Harrison and S. Johnson (1997)
Classification of the southern African sanga and east African shorthorned zebu. Anim Genet **28:** 77-83.

Fuerst, P. A., R. Chakraborty and M. Nei (1977)
Statistical studies on protein polymorphism in natural populations; I. Distribution of single locus heterozygosity. Genetics **86:** 455-483.

Gao, Q., and J. E. Womack (1997a)
Comparative mapping of anchor loci from HSA19 to cattle chromosomes 7 and 18. J Hered **88:** 524-7.

Gao, Q., and J. E. Womack (1997b)
A genetic map of bovine chromosome 7 with an interspecific hybrid backcross panel. Mamm Genome **8:** 258-61.

Georges, M., A. B. Dietz, A. Mishra, D. Nielsen, L. S. Sargeant *et al.* (1993)
Microsatellite mapping of the gene causing weaver disease in cattle will allow the study of an associated quantitative trait locus. Proc Natl Acad Sci U S A **90:** 1058-62.

Georges, M., and J. Massey (1992)
Polymorphic dna markers in bovidae (World Intellectual Property Org Geneva). WO Publ No 92/13102 .

Gillespie, J. H. (1997)
Junk ain't what junk does: neutral alleles in a selected context. Gene **205:** 291-9.

Goldammer, T., G. Guerin, R. M. Brunner, J. Vanselow, R. Furbass *et al.* (1994)
Chromosomal mapping of the bovine aromatase gene (CYP19) and an aromatase pseudogene to chromosome 10 and syntenic group U5. Mamm Genome **5:** 822-3.

Goldammer, T., R. Weikard, R. M. Brunner and M. Schwerin (1996)
Generation of chromosome fragment specific bovine DNA sequences by microdissection and DOP-PCR. Mamm Genome **7:** 291-6.

Goldstein, D. B., and A. G. Clark (1995)
Microsatellite variation in North American populations of Drosophila melanogaster. Nucleic Acids Res **23:** 3882-6.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman (1995a)
An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463-71.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza and M. W. Feldman (1995b)
Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci U S A **92:** 6723-7.

Goss, S. J., and H. Harris (1975)
New method for mapping genes in human chromosomes. Nature **255:** 680-4.

Goudet, J. (1995)
FSTAT (vers. 1.2): a computer program to calculate F-statistics. J. Hered. **86:** 485-486.

Grigson, C. (1978)
The craniology and relationships of four species of Bos. IV. The relationship between Bos primigenius Boj. and Bos taurus L. and its implications for the phylogeny of the domestic breeds. J. Arch. Sci. **5:** 123-152.

Grigson, C. (1980)
The craniology and relationships of four species of Bos. 5. Bos indicus L. J. Arch. Sci. **7:** 3-32.

Gu, Z., L. Gomez-Raya, D. I. Vage, K. Elo, W. Barendse *et al.* (2000)
Consensus and comprehensive linkage maps of bovine chromosome 7. Anim Genet **31:** 206-9.

Gu, Z., J. E. Womack and B. W. Kirkpatrick (1999)
A radiation hybrid map of bovine Chromosome 7 and comparative mapping with human Chromosome 19 p arm. Mamm Genome **10:** 1112-4.

Gu, Z. X., J. E. Womack and B. W. Kirkpatrick (1997)
Synteny mapping of four genes from the short arm of human chromosome 19 to bovine chromosome 7. Cytogenet Cell Genet **79:** 225-7.

Guerin, G., M. Nocart and S. J. Kemp (1994)
Fifteen new synteny assignments of microsatellites to the bovine genome. Anim Genet **25:** 179-81.

Guo, S. W., and E. A. Thompson (1992)
Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics **48:** 361-72.

Harr, B., B. Zangerl, G. Brem and C. Schlotterer (1998)
Conservation of locus-specific microsatellite variability across species: a comparison of two Drosophila sibling species, D. melanogaster and D. simulans. Mol Biol Evol **15:** 176-84.

Hartl, G. B., R. Göltenboth, M. Grillitsch and R. Willing (1988)
On the biochemical systematics of the Bovini. Biochem. Syst. Ecol. **16:** 575-579.

Hedrick, P. W., T. S. Whittam and P. Parham (1991)
Heterozygosity at individual amino acid sites: extremely high levels for HLA-A and -B genes. Proc Natl Acad Sci U S A **88:** 5897-901.

Heyen, D. W., J. I. Weller, M. Ron, M. Band, J. E. Beever *et al.* (1999)
A genome scan for QTL influencing milk production and health traits in dairy cattle. Physiol Genomics **1:** 165-75.

Hudson, R. R., and N. L. Kaplan (1988)
The coalescent process in models with selection and recombination. Genetics **120:** 831-40.

Hudson, R. R., and N. L. Kaplan (1995)
Deleterious background selection with recombination. Genetics **141:** 1605-17.

Hughes, A. L., and M. Nei (1988)
Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335:** 167-70.

Hughes, A. L., and M. Nei (1989)
Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci U S A **86:** 958-62.

Hughes, A. L., and M. Yeager (1997)
Comparative evolutionary rates of introns and exons in murine rodents [published erratum appears in J Mol Evol 1998 Apr;46(4):497]. J Mol Evol **45:** 125-30.

Hughes, A. L., and M. Yeager (1998)
Comparative Evolutionary Rates of Introns and Exons in Murine Rodents. J Mol Evol **46:** 497.

Jeanmougin, F., J. D. Thompson, M. Gouy, D. G. Higgins and T. J. Gibson (1998)
Multiple sequence alignment with Clustal X. Trends Biochem Sci **23:** 403-5.

Jin, L., C. Macaubas, J. Hallmayer, A. Kimura and E. Mignot (1996)
Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci U S A **93:** 15285-8.

Jones, H. B. (1996)
Hybrid selection as a method of increasing mapping power for radiation hybrids. Genome Res **6:** 761-9.

Kang'a, S., P. R. Nilsson, K. Rottengatter, T. Goldammer, C. D. Kim *et al.* (2000)
Comparative mapping of a cattle trypanotolerance QTL region on BTA7. International Society of Animal Genetics .

Kantanen, J., I. Olsaker, S. Adalsteinsson, K. Sandberg, E. Eythorsdottir *et al.* (1999)
Temporal changes in genetic variation of north European cattle breeds. Anim Genet **30:** 16-27.

Kantanen, J., I. Olsaker, L. E. Holm, S. Lien, J. Vilkki *et al.* (2000)
Genetic diversity and population structure of 20 North European cattle breeds. J Hered **91:** 446-57.

Kaplan, N. L., T. Darden and R. R. Hudson (1988)
The coalescent process in models with selection. Genetics **120:** 819-29.

Kappes, S. M., J. W. Keele, R. T. Stone, R. A. McGraw, T. S. Sonstegard *et al.* (1997)
A second-generation linkage map of the bovine genome. Genome Res **7:** 235-49.

Kaukinen, J., and S. L. Varvio (1993)
Eight polymorphic bovine microsatellites. Anim Genet **24:** 148.

Kemp, S. J., O. Hishida, J. Wambugu, A. Rink, M. L. Longeri *et al.* (1995)
A panel of polymorphic bovine, ovine and caprine microsatellite markers. Anim Genet **26:** 299-306.

Kimura, M. (1983)
The neutral theory of molecular evolution. Cambridge University Press, Cambridge .

Kimura, M., and J. F. Crow (1964)
The number of alleles that can be maintained in a finite population. Genetics **49:** 725-38.

Klein, J. (1986)
Natural history of the major histocompatibility complex. Wiley & Sons, Inc. .

Klein, J., Y. Satta, O. h. C and N. Takahata (1993)
The molecular descent of the major histocompatibility complex. Annu Rev Immunol **11:** 269-95.

Kohn, M. H., H. J. Pelz and R. K. Wayne (2000)
Natural selection mapping of the warfarin-resistance gene. Proceedings of the National Academy of Sciences of the United States of America **97:** 7911-7915.

Konfortov, B. A., V. E. Licence and J. R. Miller (1999)
Re-sequencing of DNA from a diverse panel of cattle reveals a high level of polymorphism in both intron and exon. Mamm Genome **10:** 1142-5.

Kossarek, L. M., O. Finlay, W. M. Grosse, X. Su and R. A. McGraw (1994a)
Five bovine dinucleotide repeat polymorphisms: RM026, RM029, RM032, RM033 and RM038. Anim Genet **25:** 296-7.

Kossarek, L. M., W. M. Grosse, O. Finlay and R. A. McGraw (1993)
Rapid communication: bovine dinucleotide repeat polymorphism RM006. J Anim Sci **71:** 3176.

Kossarek, L. M., W. M. Grosse, O. Finlay and R. A. McGraw (1994b)
Five bovine dinucleotide repeat polymorphisms: RM011, RM012, RM016, RM019 and RM024. Anim Genet **25:** 205-6.

Kulke, H., and D. Rothermund (1990)
A History of India. Routledge, London.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.* (2001)
Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

Lange, K., and M. Boehnke (1992)
Bayesian methods and optimal experimental design for gene mapping by radiation hybrids. Ann Hum Genet **56:** 119-44.

Lange, K., M. Boehnke, D. R. Cox and K. L. Lunetta (1995)
Statistical methods for polyploid radiation hybrid mapping. Genome Res **5:** 136-50.

Lawrence, S., N. E. Morton and D. R. Cox (1991)
Radiation hybrid mapping. Proc Natl Acad Sci U S A **88:** 7477-80.

Leach, R. J., and P. O'Connell (1995)
Mapping of mammalian genomes with radiation (Goss and Harris) hybrids. Adv Genet **33:** 63-99.

Levinson, G., and G. A. Gutman (1987)
Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol **4:** 203-21.

Li, Y., T. Fahima, J. H. Peng, M. S. Roder, V. Kirzhner *et al.* (2000a)
Edaphic microsatellite DNA divergence in wild emmer wheat, *Triticum dicoccoides*, at a microsite: Tabigha, Israel. Theor Appl Genet **101:** 1029-1038.

Li, Y., M. S. Roder, T. Fahima, V. Kirzhner, A. Beiles *et al.* (2000b)
Natural selection causing microsatellite divergence in wild emmer wheat at the ecologically variable microsite at Ammiad, Israel. Theor Appl Genet **100:** 985-999.

Lichten, M., and A. S. Goldman (1995)
Meiotic recombination hotspots. Annu Rev Genet **29:** 423-44.

Loftus, R. T., O. Ertugrul, A. H. Harba, M. A. El-Barody, D. E. MacHugh *et al.* (1999)
A microsatellite survey of cattle from a centre of origin: the Near East. Mol Ecol **8:** 2015-22.

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:** 2757-61.

Louis, E. J., and E. R. Dempster (1987)
An exact test for Hardy-Weinberg and multiple alleles. Biometrics **43:** 805-11.

Lunetta, K. L., and M. Boehnke (1994)
Multipoint radiation hybrid mapping: comparison of methods, sample size requirements, and optimal study characteristics. Genomics **21:** 92-103.

Lunetta, K. L., M. Boehnke, K. Lange and D. R. Cox (1996)
Selected locus and multiple panel models for radiation hybrid mapping. Am J Hum Genet **59:** 717-25.

Lyons, L. A., T. F. Laughlin, N. G. Copeland, N. A. Jenkins, J. E. Womack *et al.* (1997)
Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nat Genet **15:** 47-56.

Ma, R. Z., M. J. van Eijk, J. E. Beever, G. Guerin, C. L. Mummery *et al.* (1998)
Comparative analysis of 82 expressed sequence tags from a cattle ovary cDNA library. Mamm Genome **9:** 545-9.

MacHugh, D. E., (1996)
Molecular Biogeography and Genetic Structure of Domesticated Cattle, Department of Genetics. University of Dublin, Dublin.

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:** 1071-86.

Maré, C. J. (1988)
African Animal Trypanosomiasis, in *Foreign Animal Diseases (6th ed.)*. Committee on Foreign Animal Diseases of the United States Animal Health Association.

McGraw, R. A., W. M. Grosse, S. M. Kappes, C. W. Beattie and R. T. Stone (1997)
Thirty-four bovine microsatellite markers. Anim Genet **28:** 66-8.

Meadow, R. H. (1984)
Animal domestication in the Middle East: a view from the Eastern margin., pp. 309-337 in *Animals and Archaeology: 3. Early Herders and their Flocks*, edited by J. Clutton-Brock and C. Grigson. Oxford University Press, Oxford.

Meadow, R. H. (1993)
Animal domestication in the Middle East: A revised view from the Eastern Margin., pp. 295-320 in *Harappan Civilisation*, edited by G. Possehl. Oxford & IBH., New Delhi.

Meagher, S., and W. K. Potts (1997)
A microsatellite-based MHC genotyping system for house mice (Mus domesticus). Hereditas **127:** 75-82.

Mezzelani, A., Y. Zhang, L. Redaelli, B. Castiglioni, P. Leone *et al.* (1995)
Chromosomal localization and molecular characterization of 53 cosmid-derived bovine microsatellites. Mamm Genome **6:** 629-35.

Minch, E. (1995)
Microsat ver. 1.5d. .

Moore, S. S., W. Barendse, K. T. Berger, S. M. Armitage and D. J. Hetzel (1992)
Bovine and ovine DNA microsatellites from the EMBL and GENBANK databases. Anim Genet **23:** 463-7.

Moore, S. S., and K. Byrne (1993)
Dinucleotide polymorphism at the bovine histamine H1 receptor locus. Anim Genet **24:** 72.

Moore, S. S., K. Byrne, K. T. Berger, W. Barendse, F. McCarthy *et al.* (1994)
Characterization of 65 bovine microsatellites. Mamm Genome **5:** 84-90.

Morkos, N. B. B., M. D. Grosz and R. T. Stone (1994)
Placement of BoLA-DIB into a microsatellite-based linkage group. Anim Genet (Suppl 2) **25:** 55.

Murray, M., W. I. Morrison and D. D. Whitelaw (1982)
Host susceptibility to African trypanosomiasis: trypanotolerance. Adv Parasitol **21:** 1-68.

Nei, M. (1972)
Genetic distance between populations. American Naturalist **106:** 283-92.

Nei, M. (1973)
Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A **70:** 3321-3.

Nei, M. (1987)
Molecular Evolutionary Genetics. Columbia University Press, New York, NY, USA.

Nei, M., T. Maruyama and R. Chakraborty (1975)
The bottleneck effect and genetic variability in populations. Evolution **29:** 1-10.

Nickerson, D. A., S. L. Taylor, K. M. Weiss, A. G. Clark, R. G. Hutchinson *et al.* (1998)
DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nat Genet **19:** 233-40.

Nordborg, M., B. Charlesworth and D. Charlesworth (1996)
The effect of recombination on background selection. Genet Res **67:** 159-74.

O'Brien, S. J., J. E. Womack, L. A. Lyons, K. J. Moore, N. A. Jenkins *et al.* (1993)
Anchored reference loci for comparative genome mapping in mammals. Nat Genet **3:** 103-12.

Ohta, T., and M. Kimura (1973)
A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetical Research **22:** 210-4.

Ota, T., (1993)
DISPAN: Genetic distance and phylogenetic analysis., Pennsylvania State University, USA.

Ozawa, A., M. R. Band, J. H. Larson, J. Donovan, C. A. Green *et al.* (2000)
Comparative organization of cattle chromosome 5 revealed by comparative mapping by annotation and sequence similarity and radiation hybrid mapping. Proc Natl Acad Sci U S A **97:** 4150-5.

Paterson, S. (1998)
Evidence for balancing selection at the major histocompatibility complex in a free-living ruminant. J Hered **89:** 289-94.

Payne, W. J. A. (1970)
Cattle Production in the Tropics. Longman, London.

Perkins, D., (1969)
Fauna of Catal Huyuk: evidence for early cattle domestication in Anatolia. Science **164:** 177-9.

Raeymaekers, P., K. Van Zand, L. Jun, M. Hoglund, J. J. Cassiman *et al.* (1995)
A radiation hybrid map with 60 loci covering the entire short arm of chromosome 12. Genomics **29:** 170-8.

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

Rexroad, C. E., 3rd, E. K. Owens, J. S. Johnson and J. E. Womack (2000)
A 12,000 rad whole genome radiation hybrid panel for high resolution mapping in cattle: characterization of the centromeric end of chromosome 1. Anim Genet **31:** 262-5.

Rexroad, C. E., J. S. Schlapfer, Y. Yang, B. Harlizius and J. E. Womack (1999)
A radiation hybrid map of bovine chromosome one. Anim Genet **30:** 325-32.

Rice, W. R. (1989)
Analyzing tables of statisical tests. Evolution **43:** 223-5.

Ruvolo, M. (1997)
Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol Biol Evol **14:** 248-65.

Schlapfer, J., Y. Yang, C. Rexroad, 3rd and J. E. Womack (1997)
A radiation hybrid framework map of bovine chromosome 13. Chromosome Res **5:** 511-9.

Schlotterer, C., R. Ritter, B. Harr and G. Brem (1998)
High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. Mol Biol Evol **15:** 1269-74.

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

Schneider, S., J. Kueffer, D. Roessli and L. Excoffier (1997)
Arlequin ver. 1.1: Software for population genetic data analysis. .

Sherry, S. T., H. C. Harpending, M. A. Batzer and M. Stoneking (1997)
Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics **147:** 1977-82.

Skow, L. C., J. Goy and D. Honeycutt (1994)
Dinucleotide repeat polymorphism near a bovine MHC class I sequence. Anim Genet **25:** 290.

Slatkin, M. (1994)
An exact test for neutrality based on the Ewens sampling distribution. Genet Res **64:** 71-4.

Slatkin, M. (1995a)
Hitchhiking and associative overdominance at a microsatellite locus. Mol Biol Evol **12:** 473-80.

Slatkin, M. (1995b)
A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457-62.

Slatkin, M., and T. Wiehe (1998)
Genetic hitch-hiking in a subdivided population. Genet Res **71:** 155-60.

Smit, A. G., P,
RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html. .

Smith, J. M., and J. Haigh (1974)
The hitch-hiking effect of a favourable gene. Genet Res **23:** 23-35.

Smith, T. P., N. Lopez-Corrales, M. D. Grosz, C. W. Beattie and S. M. Kappes (1997)
Anchoring of bovine chromosomes 4, 6, 7, 10, and 14 linkage group telomeric ends via FISH analysis of lambda clones. Mamm Genome **8:** 333-6.

Solinas-Toldo, S., C. Lengauer and R. Fries (1995)
Comparative genome map of human and cattle. Genomics **27:** 489-96.

Steffen, P., A. Eggen, A. B. Dietz, J. E. Womack, G. Stranzinger *et al.* (1993)
Isolation and mapping of polymorphic microsatellites in cattle. Anim Genet **24:** 121-4.

Stone, R. T., and N. E. Muggli-Cockett (1993)
BoLA-DIB: species distribution, linkage with DOB, and northern analysis. Anim Genet **24:** 41-5.

Stone, R. T., J. C. Pulido, G. M. Duyk, S. M. Kappes, J. W. Keele *et al.* (1995)
A small-insert bovine genomic library highly enriched for microsatellite repeat sequences. Mamm Genome **6:** 714-24.

Sun, H. S., G. L. Hart and B. W. Kirkpatrick (1993a)
A polymorphic microsatellite (UWCA1) detected on bovine chromosome 23. Anim Genet **24:** 142.

Sun, H. S., S. Whallon, A. Ponce De Leon and B. W. Kirkpatrick (1993b)
Development of polymorphic bovine microsatellite markers from a cosmid library. J Anim Sci (Suppl) **71:** 100.

Takezaki, N., and M. Nei (1996)
Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics **144:** 389-99.

Taylor, M. F. J., Y. Shen and M. E. Kreitman (1995)
A Population Genetic Test of Selection At the Molecular-Level. Science **270:** 1497-1499.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994)
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673-80.

Toldo, S. S., R. Fries, P. Steffen, H. L. Neibergs, W. Barendse *et al.* (1993)
Physically mapped, cosmid-derived microsatellite markers as anchor loci on bovine chromosomes. Mamm Genome **4:** 720-7.

Trail, J. C. M., C. H. Hoste, Y. J. Wissocq and P. Lhoste, 1979 Trypanotolerant livestock in West and Central Africa, International Livestock Centre for Africa, Addis Ababa, Ethiopia.

Troy, C. S., D. E. MacHugh, J. F. Bailey, D. A. Magee, R. T. Loftus *et al.* (2001)
Genetic evidence for Near-Eastern origins of European cattle. Nature **410:** 1088-91.

Turpeinen, T., T. Tenhola, O. Manninen, E. Nevo and E. Nissila (2001)
Microsatellite diversity associated with ecological factors in Hordeum spontaneum populations in Israel. Mol Ecol **10:** 1577-91.

Vaiman, D., D. Mercier, K. Moazami-Goudarzi, A. Eggen, R. Ciampolini *et al.* (1994)
A set of 99 cattle microsatellites: characterization, synteny mapping and polymorphism. Mamm Genome **5:** 288-97.

Vaiman, D., R. Osta, D. Mercier, C. Grohs and H. Leveziel (1992)
Characterization of five new bovine dinueleotide repeats. Anim Genet **23:** 537-41.

Valdes, A. M., M. Slatkin and N. B. Freimer (1993)
Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737-49.

van Eijk, M. J., J. E. Beever, Y. Da, J. A. Stewart, G. E. Nicholaides *et al.* (1995)
Genetic mapping of BoLA-A, CYP21, DRB3, DYA, and PRL on BTA23. Mamm Genome **6:** 151-2.

van Haeringen, W. A., P. S. Gwakisa, S. Mikko, E. Eythorsdottir, L. E. Holm *et al.* (1999)
Heterozygosity excess at the cattle DRB locus revealed by large scale genotyping of two closely linked microsatellites. Anim Genet **30:** 169-76.

Walter, M. A., D. J. Spillett, P. Thomas, J. Weissenbach and P. N. Goodfellow (1994)
A method for constructing radiation hybrid maps of whole genomes. Nat Genet **7:** 22-8.

Watterson, G. A. (1977)
Heterosis or neutrality? Genetics **85:** 789-814.

Watterson, G. A. (1978)
The homozygosity test of neutrality. Genetics **88:** 405-417.

Weber, J. L., and C. Wong (1993)
Mutation of human short tandem repeats. Hum Mol Genet **2:** 1123-8.

Weikard, R., T. Goldammer, C. Kuhn, W. Barendse and M. Schwerin (1997)
Targeted development of microsatellite markers from the defined region of bovine chromosome 6q21-31. Mamm Genome **8:** 836-40.

Weir, B. S., and C. C. Cockerham (1984)
Estimating F-statistics for the analysis of population structure. Evolution **38:** 1358-70.

Wendorf, F., and R. Schild (1994)
Are the early Holocene cattle in the Eastern Sahara domestic or wild? Evol. Anthropol. **3.**

Wiehe, T. (1998)
The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. Theor Popul Biol **53:** 272-83.

Womack, J. E., J. S. Johnson, E. K. Owens, C. E. Rexroad, 3rd, J. Schlapfer *et al.* (1997)
A whole-genome radiation hybrid panel for bovine gene mapping. Mamm Genome **8:** 854-6.

Womack, J. E., and Y. D. Moll (1986)
Gene map of the cow: conservation of linkage with mouse and man. J Hered **77:** 2-7.

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

Yang, Y. P., C. E. Rexroad, 3rd, J. Schlapfer and J. E. Womack (1998)
An integrated radiation hybrid map of bovine chromosome 19 and ordered comparative mapping with human chromosome 17. Genomics **48:** 93-9.

Zhang, N., and J. E. Womack (1992)
Synteny mapping in the bovine: genes from human chromosome 5. Genomics **14:** 126-30.