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

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

Copyright statement

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

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

Liability statement

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

Access Agreement

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

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

GENETIC ADMIXTURE ANALYSIS OF DOMESTICATED CATTLE

Abigail Ruth Freeman

A thesis submitted to the University of Dublin

for the degree of Doctor of Philosophy

Department of Genetics

University of Dublin

Trinity College

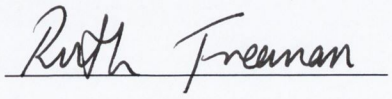
April 2003

TRINITY COLLEGE
0 4 NOV 2003
LIBRARY DUBLIN
THOSIS
7511
7379

DECLARATION

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

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



Abigail Ruth Freeman

April 2003

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible. I have to begin with my family, who are a perpetual source of encouragement and always believed that I could do this. Mum, Dad, Jane and Suzy…I'd be lost without you all, and apologise if I was intolerable during the writing of this thesis! I also want to thank my grandmothers, Grace and Margaret, who think I'm great, no matter what I do.

My supervisor, Dan Bradley, has taught me a great deal and been a constant source of new ideas. I want to thank him for careful guidance throughout the years and for giving me the opportunity to work on this project. I am especially grateful for having been given the chance to do some of the work in Nairobi. He made me study statistics, but in return I inflicted the "payneful" task of proofreading this thesis! He has turned a blind eye to my fruitless searches for Asiatic cheetah samples, when I probably should have been doing something else. One simply could not wish for a better or friendlier supervisor.

I want to say a huge thank you to Olivier Hanotte, John Gibson and everyone in Lab 7 at ILRI, Nairobi, Kenya. It was a wonderful experience to work in such a diverse environment, and collaborate with the talented people at ILRI. All of the technicians in Lab 7 provided able assistance, in particular Paul Wafula and Joel Ochieng. My fellow hosteller, Liam Morrison, "dragged" me to the ILRI bar on many an evening, and was a much-appreciated lunchtime curry and shopping companion when compound living became a bit claustrophobic.

Data presented in chapters 2 and 3 of this thesis were produced by Pushpendra Kumar and Ciaran Meghan and I am very grateful for their permission to include it. Thanks also to Teresa Cymbron for permitting some of her Mediterranean data to be included in the analysis in Chapter 5. In my daily work I have been blessed with a friendly and obliging group of co-workers. Thanks to all the original cowboys (and girls); Ronan, Dave, John, Jill, Ashie, Emmy and Chris, for showing me the ropes and introducing me to long liquid nights at La Cave. I couldn't have shared a lab bench with a more agreeable person than Dave Magee, and really I enjoyed most of the football commentary. Stephen, whose endless knowledge and talents constantly astound me, and whose programming skills have reduced the workload of microsatters everywhere, deserves much thanks for all his assistance. Thankfully, Ceiridwen is as bad as me when it comes to insisting that the lab is kept tidy, so I thank her for making me feel slightly less neurotic! I wish Brian, Laoise and Emma all the very best in getting finished.

Without the valuable technical and administrative support of the Genetics Department staff many of the studies reported in this thesis simply would not have been possible. So a special mention must go to the ever-cheerful Dave "stop chatting and get back to work" Sullivan, Paul, Brenda, Louis, Sue, Kay, and Vera. It would be a long list to mention all the other member of the Genetics department who have helped me out over the years but I gratefully thank all of them. I have to single out Ashie for getting me running a few years ago, and Louise for encouraging me to run further than I thought humanly possible! Thanks girls. Also, the Identigirls, who have always provided plenty of chat, laughs and chocolate. I also owe a lot to Simon Wong and Jonathan McCrea, who helped to avoid PC disasters of Gatesian proportions.

Life would have been a little quieter and much less interesting during this project without my invaluable friends. While I suspect that most of them have no idea what I have been doing for the last few years, this doesn't stop them being amazingly encouraging. Scott, Kate, Johnny, Graham, Dara Lynne, Tina, Guy, Mark, Sarah B, Tanya, Becs, Sophie, Ollie, Solveig, Sarah N, Judith, Charlie, Pete, Dec and Ivan, thanks for being there through good times and bad. And no, I have not been developing either inordinately large tomatoes, exceptionally small dogs, or keeping either cheetahs or chimps called cheetah locked away in cages! A special mention goes to Dave MacHugh, who has always been an extraordinary source of advice on all things genetic. Despite being a bit of a know-it all (made all the more annoying by the fact that he does!), he helped me realise the importance of a sound working method in doing research and of avoiding MACDRAW at all costs. Thanks for everything, especially for dealing with the first year Ph.D. blues, and for proofreading parts of this thesis.

During this project I have had the chance to interact with lots of people, so I have invariably forgotten to thank someone. If I have failed to mention your name, this does not mean I am less grateful. Finally, I want to dedicate this thesis to my parents, David and Valerie. I really couldn't have done it without their support and encouragement. They have taught me that anything can be achieved with dedication, and regularly demonstrate this by example. Thank you for everything.

CONTENTS

DECLARATION ii
ACKNOWLEDGEMENTS iii
CONTENTS v
FIGURES x
TABLES xiv
ABBREVIATIONS xv
SUMMARY xvii

CHAPTER 1. GENERAL INTRODUCTION 1

1.1 INTRODUCTION 2

1.2 CATTLE 5

1.2.1 Domestication of cattle 5

1.2.2 Migration of cattle from original domestication centres 7
1.2.2.a Migration to Europe 7
1.2.2.b Migration to Africa 7
1.2.2.c Migration to India and the Near East 10

1.3 LINKAGE DISEQUILIBRIUM 12

1.4 ADMIXTURE 16

1.4.1 A definition of admixture 16

1.4.2 Measuring admixture 16
1.4.2.a Methods of admixture estimation 16
1.4.2.b Population-associated alleles and admixture estimation 19
1.4.2.c Bayesian methods to estimate admixture 21
1.4.2.d Other methods of admixture estimation 26

1.4.3 Why the interest in admixed populations? 27

1.5 MAPPING BY ADMIXTURE GENERATED LINKAGE DISEQUILIBRIUM 28

1.6 CATTLE AND TRYPANOSOMIASIS 31

1.7 MICROSATELLITES 34

1.7.1 Introduction to microsatellites 34

1.7.2 Microsatellite mutation 34
1.7.2.a Infinite allele model (IAM) 35
1.7.2.b Stepwise mutation model (SMM) 35
1.7.2.c Two phase model 35

1.8 OVERVIEW 37

CHAPTER 2. ADMIXTURE ANALYSIS OF SOUTH ASIAN CATTLE 38

2.1 INTRODUCTION 39

2.2 MATERIALS AND METHODS 43

2.2.1 Samples 43
2.2.1.a Hariana 43
2.2.1.b Sahiwal 43
2.2.1 c Red Sindhi 44
2.2.1.d Tharparker 44
2.2.1.e Desi 44
2.2.1.f Nellore 44
2.2.1.g Ongole 45

2.2.2 Genotyping 45

2.2.3 Analysis of microsatellite data 46
2.2.3.a Mean Number of Alleles 46
2.2.3.b Gene Diversity 47
2.2.3.c Population-Associated Alleles 48
2.2.3.d Genetic distance 50
2.2.3.e STRUCTURE analysis 52

2.3 RESULTS 53

2.3.1 Diversity measures 53

2.3.2 Selection of Population Associated Alleles (PAAs) 53
2.3.2.a Support for PAA selection (1) 56
2.3.2.b Support for PAA selection (2) 57

2.3.3 Genetic Distance 58
2.3.3.a Application of PAAs to the Near East 58
2.3.3.b Application of PAAs to India 58

2.3.4 STRUCTURE analysis 60

2.4 DISCUSSION 62

2.4.1 Patterns of genetic variability in the breeds surveyed 62

2.4.2 Introgression of *B. taurus* 62

CHAPTER 3. A POPULATION GENETIC SURVEY OF WEST AFRICAN CATTLE POPULATIONS 68

3.1 INTRODUCTION 69

3.2 MATERIALS AND METHODS 72

3.2.1 Sample Collection and Genotyping 72

3.2.2 Diversity measures 75

3.2.3 Admixture analysis 75

3.3 RESULTS 80

3.3.1 Diversity measures 80

3.3.2 Multidimensional scaling analysis. 80

3.3.3 Genetic distance 81

3.4.4 Admixture 84
3.4.4.a Introgression of Bos indicus *85*
3.4.4.b Introgression of European/Near Eastern B. taurus *93*
3.4.4.c African B. taurus – the genetic background. *96*

3.5 DISCUSSION 98

3.5.1 Genetic diversity and admixture 98

3.5.2 Correspondence of admixture measures. 102

3.5.3 Application of admixture studies to West Africa 104

CHAPTER 4. HAPLOTYPE BREAKDOWN IN THE BOVINE HYBRID ZONES OF AFRICA AND THE NEAR EAST: linkage disequilibrium and applications to admixture mapping. 106

4.1 INTRODUCTION 107

4.1.1 Cattle in eastern and southern Africa 107

4.1.2 Hybrid populations and haplotype breakdown 110

4.1.3 Admixture detection 113
4.1.3.a Mismatch analysis *113*
4.1.3.b Linkage disequilibrium *115*
4.1.3.c Allele ancestry *115*

4.2 MATERIALS AND METHODS 116

4.2.1 Sample information 116

4.2.2 Selection of X-chromosome microsatellites 118

4.2.3 Genotyping of 15 X-chromosome microsatellites 121
4.2.3.a Radioactive PCR amplification *121*
4.2.3.b Polyacrylamide Gel Electrophoresis (PAGE) of radioactive PCR product *124*

4.2.4 Genotyping of 10 X-chromosome microsatellites in a subset of samples 125
4.2.4.a PCR amplification *125*
4.2.4.b PAGE of PCR products *127*

4.2.5 SNP detection and genotyping **127**
4.2.5.a Amplicon selection and primer design *127*
4.2.5.b PCR amplification of intronic sections *128*
4.2.5.c SNP detection and Single Strand Conformational Polymorphism (SSCP) primer design and PCR amplification *131*
4.2.5.d Gel Electrophoresis of PCR-SSCP products *133*

4.2.6 Analysis of microsatellite data **135**
4.2.6.a Allele frequencies *135*
4.2.6.b Phylogeography *135*
4.2.6.c Mismatch distributions *135*
4.2.6.d Linkage Disequilibrium *136*
4.2.6.e Allele ancestry *137*

4.2.7 Analysis of SNP data **141**
4.2.7.a Allele frequencies *141*
4.2.7.b Allele ancestry *141*

4.3 RESULTS **143**

4.3.1 RESULTS 1 **143**
Dataset 1 – 15 markers in 13 breeds *143*
4.3.1.a X- chromosome maps *143*
4.3.1.b Allele frequencies *144*
4.3.1.c Phylogeography *144*
4.3.1.d Mismatch Distributions *147*
4.3.1.e Linkage disequilibrium between linked markers *149*
4.3.1.f Linkage disequilibrium between linked and non-linked markers *150*
4.3.1.g Fully and Partially Bayesian Analysis of Allele Ancestry. *152*
4.3.1.h Estimation of number of recombinations *158*

4.3.2 RESULTS 2 **161**
Dataset 2 – 10 markers in twenty-two breeds *161*
4.3.2.a Allele frequencies *161*
4.3.2.b Phylogeography *161*
4.3.2.c Mismatch analysis *164*
4.3.2.d Linkage disequilibrium *164*
4.3.2.e Fully and Partially Bayesian Analysis of Allele Ancestry. *167*
4.3.2.f Estimation of number of recombinations *171*

4.3.3 RESULTS 3 **175**
SNP analysis *175*
4.3.3.a Allele frequencies *175*
4.3.3.b Allele ancestry *175*
4.3.3.c Estimation of number of recombinations *178*

4.4 DISCUSSION **180**

4.4.1 Mismatch Distribution **182**

4.4.2 Linkage Disequilibrium **184**
4.4.2.a Linkage disequilibrium and mapping *187*
4.4.2.b LD as a tool for evolutionary inference *190*

4.4.3 Allele ancestry and recombination **192**

4.4.4 Applications and Future Work **200**

CHAPTER 5. GENETIC ANALYSIS USING MULTIPLE MICROSATELLITE DATASETS 202

5.1 INTRODUCTION 203

5.1.1 Data 203

5.1.2 Bayesian analysis of admixture 205

5.1.3 Combining data 207

5.2 MATERIALS AND METHODS 209

5.2.1 Samples 209

5.2.2 Calculation of admixture for multiple values of k. 212

5.2.3 Combination of Diversity and Admixture Measures 212
5.2.3.a Calculation of variables for combination *212*
5.2.3.b Combination of variables *214*
5.2.3.c Visualisation of transformed combined data. *216*

5.3 RESULTS 218

5.3.1 Population structure within datasets 218
5.3.1.a Dataset 1 *219*
5.3.1.b Dataset 2 *223*
5.3.1.c Dataset 3 *226*
5.3.1.d High values of k *231*

5.3.2 Combined Datasets 235
5.3.2.a Combining Data *235*
5.3.2.b Unbiased Heterozygosity ($\hat{H}$) *241*
5.3.2.c Mean number of alleles ($\hat{A}$) *243*
5.3.2.d Admixture *246*

5.4 DISCUSSION 251

5.4.1 Population structure 251

5.4.2 Global Diversity 258

5.4.3 Methodological issues. 262

CHAPTER 6. REFERENCES 265

APPENDIX A Microsatellite allele frequencies 279

APPENDIX B Breed codes 310

APPENDIX C Solutions 312

FIGURES

Chapter 1

Figure 1.1 *Map showing domestication centres and subsequent routes of domestic cattle.* **10**

Figure 1.2 *Distribution of modern domestic cattle types.* **11**

Figure 1.3 *Hybrid isolation (HI) and Continual Gene Flow (CGF) models of linkage disequilibrium decay.* **14**

Figure 1.4 *Monte Carlo simulation, as applied to a biological system.* **23**

Figure 1.5 *Schematic illustrating the principles of mapping by admixture-generated linkage disequilibrium.* **29**

Figure 1.6 *Tsetse fly on human skin.* **31**

Figure 1.7 *Trypanosomes (T. brucei) parasites.* **32**

Chapter 2

Figure 2.1 *Map of the Indo-Pakistan region showing the site of origin of six Indian cattle breeds.* **46**

Figure 2.2 *Histogram of Scaled Frequency Differential (SFD) values.* **55**

Figure 2.3 *Histogram of Allele Adjacency Index (AAI) for 20 simulated microsatellite loci.* **57**

Figure 2.4 *Scatter plots of average of Nei's standard genetic distance from Europe vs. the average frequency of PAAs.* **59**

Figure 2.5 *Box plot of mean individual admixture coefficient ($\widehat{q}$) for each individual grouped by breed.* **61**

Chapter 3

Figure 3.1 *Map showing sampling sites of West African cattle breeds.* **73**

Figure 3.2 *Contour map of size adjusted mean number of alleles.* **82**

Figure 3.3 *Multi-dimensional scaling plot of F_{ST} values.* **83**

Figure 3.4 *Neighbor-joining tree constructed using a matrix of Nei's standard genetic distances.* **84**

Figure 3.5 *Histograms of six admixture measures of Bos indicus introgression in African cattle breeds.* **88**

Figure 3.6 *Line plot of log likelihoods of different values of k.* **91**

Figure 3.7 *Contour map of mean $\boldsymbol{q}_i$, the admixture coefficient characterised by Indian Bos indicus.* **92**

Figure 3.8 *Contour map of a normalised value of $\boldsymbol{q}_e$ the admixture coefficient*

characterised by European/Near Eastern Bos taurus. 95

Figure 3.9 *Regression scatter plot of the second principal component and the normalised values of* q_e. 96

Figure 3.10 *Contour map of* q_a, *the admixture coefficient characterised by African Bos taurus* 97

Chapter 4

Figure 4.1 *Regression of second principal component vs. distance from putative origin of African pastoralism.* 109

Figure 4.2 *Schematic of haplotype breakdown over time.* 112

Figure 4.3 *Hypothetical mismatch distributions of X-chromosome haplotypes following hybridisation events.* 114

Figure 4.4 *Map of approximate sampling sites for cattle populations analysed in Chapter 4.* 118

Figure 4.5 *Radiation hybrid and linkage maps of the bovine X-chromosome.* 120

Figure 4.6 *Schematic showing relative order of relevant markers on BTAX.* 120

Figure 4.7 *SSCP band shift on polyacryamide gel visualised by phosphoimager.* 134

Figure 4.8 *Multi-dimensional scaling plot of population pairwise* F_{ST} *values calculated from 15 X-chromosome microsatellites.* 145

Figure 4.9 *Unrooted neighbour-joining tree of individual X-chromosome haplotypes constructed from alleles sharing differences between 15 microsatellites for 240 individuals.* 146

Figure 4.10 *Mismatch distributions calculated from 15 marker haplotypes using the two models of microsatellite evolution.* 148

Figure 4.11 *Pairwise linkage disequilibrium between 15 linked X-chromosome microsatellite markers.* 150

Figure 4.12 *Pairwise linkage disequilibrium results for linked and non-linked markers.* 151

Figure 4.13 *X-chromosome haplotypes reconstructed from* STRUCTURE *fully Bayesian allele designation.* 155

Figure 4.14 *X-chromosome haplotypes reconstructed from a partially Bayesian analysis of allele ancestry.* 157

Figure 4.15 *Histogram of average number of recombinations in the surveyed region of the X-chromosome based on probable allele ancestry.* 158

Figure 4.16 *Linear regression plot showing average number of recombinations detected by partially Bayesian methods plotted against admixture coefficient (mR).* 159

Figure 4.17 *Multidimensional scaling plot of population pairwise* F_{ST} *values for 22 cattle breeds, genotyped for 10 microsatellite markers.* 162

Figure 4.18 *Unrooted neighbor-joining tree showing individual X-chromosome haplotypes constructed from 10 linked microsatellite markers.* 163

Figure 4.19 *Mismatch distributions constructed from 10 microsatellite markers using the two models of microsatellite evolution.* 165

Figure 4.20 *Linkage disequilibrium results for Southern and Eastern cattle populations genotyped with 10 linked and 15 unlinked microsatellite markers.* **166**

Figure 4.21 *Fully Bayesian analysis of allele ancestry for 10 microsatellite markers according to their mapped order on the X-chromosome.* **168**

Figure 4.22 *Partial Bayesian allele ancestry assignment for ten microsatellite markers according to their mapped order on the X-chromosome.* **170**

Figure 4.23 *Histogram of recombination frequency in the surveyed region of the X-chromosome based on probable allele ancestry.* **172**

Figure 4.23 *Scatter plot of regression between admixture coeffiecient and average number of recombinations detected by partially Bayesian allele ancestry model.* **173**

Figure 4.24 *Map of residual values for 22 cattle breeds.* **174**

Figure 4.25 *Allele frequencies of novel SNPs discovered in intron 8 of the bovine ectodysplasin gene.* **176**

Figure 4.26 *Bayesian and partially Bayesian analysis of three X-chromosome SNPs and a closely linked microsatellite.* **177**

Figure 4.26 *Scatter plot of regression between admixture coeffiecient and average number of recombinations detected by partially Bayesian allele ancestry model between four physically linked X-chromosome markers.* **179**

Figure 4.27 *Cumulative frequency plots of the significance of LD between markers genotyped in an admixed sheep population.* **186**

Figure 4.28 *Regression of the first principal component 1 of Hanotte et al, 2002, vs. distance from the Horn of Africa.* **194**

Figure 4.29 *Map of Africa showing the sub-Saharan Sahel region.* **196**

Chapter 5

Figure 5.1 *Plot of log likelihood values of different k values for each of the three surveyed datasets.* **219**

Figure 5.2 *DISTRUCT plot showing individual admixture levels from STRUCTURE analysis of Dataset 1 with the number of inferred populations set from two to seven.* **221**

Figure 5.3 *DISTRUCT plot showing breed admixture levels from STRUCTURE analysis of Dataset 1 with the number of inferred populations set from two to seven.* **222**

Figure 5.4 *DISTRUCT plot showing individual admixture levels from STRUCTURE analysis of Dataset 2 with the number of inferred populations set from two to seven.* **224**

Figure 5.5 *DISTRUCT plot showing breed admixture levels from STRUCTURE analysis of Dataset 2 with the number of inferred populations set from two to seven.* **225**

Figure 5.6 *DISTRUCT plot showing individual admixture levels from STRUCTURE analysis of Dataset 3 with the number of inferred populations set from two to seven.* **229**

Figure 5.7 *DISTRUCT plot showing breed admixture levels from STRUCTURE analysis of Dataset 3 with the number of inferred populations set from two to seven.* **230**

Figure 5.8 *DISTRUCT diagram of 3 datasets with k set to 15, 20 and 20.* **234**

Figure 5.9 *Schematic showing the average line produced from the regression of variable **a** on variable **b**, and the reciprocal of the regression of variable **b** on variable **a**.* **236**

Figure 5.10 *Regression plots of q_i values for common breeds found in Datasets 1 and 4.* **237**

Figure 5.11 *Regression plots of q_i values for common breeds found in Pred_Dataset1_q_i and Pred_Dataset4_ q_i.* **238**

Figure 5.12 *Scatterplots of regressions carried out between raw and transformed datasets* **239**

Figure 5.13 *Contour map of average predicted $\hat{H}$.* **242**

Figure 5. 14 *Contour map of size-adjusted mean number of alleles.* **245**

Figure 5.15 *Scatterplots showing the relationship between $\bar{A}$ and $\hat{H}$ in Datasets 1, 2 and 3 and combined datasets.* **246**

Figure 5.16 *Contour map of new predicted q_a* **248**

Figure 5.17 *Contour map of new predicted q_e* **249**

Figure 5.18 *Contour map of new predicted q_L* **250**

TABLES

Table 2.1	*Table of breed names, codes and diversity values.*	**53**
Table 3.1	*Cattle breeds included in the survey presented in Chapter 3.*	**74**
Table 3.2	*Table of correlations for measures of Bos indicus admixture.*	**86**
Table 4.1	*Cattle breeds included in the survey presented in Chapter 4.*	**117**
Table 4.2	*Primer sequences and optimal PCR conditions for amplification of 15 X-chromosome microsatellites.*	**122**
Table 4.3	*Table of PCR components for amplification of X-chromosome microsatellites incorporating a radioactive label.*	**124**
Table 4.4	*Table of optimal PCR conditions and fluorescent dye labels used in the amplification of samples for analysis on an ABI377 automated DNA sequencer.*	**126**
Table 4.5	*Table of primer sequences to amplify approximately 1kb from 4 of the introns of the bovine ED1 gene.*	**128**
Table 4.6	*Optimal PCR conditions for amplification of 1kb sections of 4 introns from the ED1 gene of B. taurus.*	**129**
Table 4.7	*Table of PCR components for amplification of 1kb section of introns of ED1 gene of B. taurus.*	**130**
Table 4.8	*Table of primer sequences to amplify approximately 300-500 bp of intron 8 of the bovine ED1 gene.*	**131**
Table 4.9	*Table of PCR components for amplification of 300-500 bp sections of intron 8 of ED1 gene of B. taurus.*	**133**
Table 4.10	*Optimal PCR conditions amplification of 300-500 bp sections of intron 8 of ED1 gene of B. taurus*	**133**
Table 5.1	*List of breed population sizes (N) and average predicted diversity ($\hat{H}$ and $\bar{A}$) and admixture variables (q_e, q_i and q_a).*	**210**
Table 5.2a	*Interpretation of clusters produced by Bayesian admixture analysis of Dataset 1.*	**220**
Table 5.2b	*Interpretation of clusters produced by Bayesian admixture analysis of Dataset 2.*	**226**
Table 5.2c	*Interpretation of clusters produced by Bayesian admixture analysis of Dataset 3.*	**228**
Table 5.3	*Regression parameters for transformation of diversity parameters.*	**243**
Table 5.4	*Regression parameters for pairwise regression analyses of admixture coefficients.*	**247**

ABBREVIATIONS

$\hat{q}$	Mean of the posterior distribution of admixture coefficient estimate by STRUCTURE
$\hat{h}$	Nei's unbiased heterozygosity for one locus
$\hat{H}$	Nei's unbiased heterozygosity for a multilocus system
δ	delta – frequency differential of a marker or trait of interest between two populations
µCi	microcurie
µl	microlitre
µM	micromolar
$\bar{A}$	Mean number of alleles
AAI	Allele adjacency index
bp	base pairs
BP	Before present
BTAX	Bovine X chromosome
c.	circa
cM	centimorgans
ddH_20	Deionised distilled water
DNA	Deoxyribose nucleic acid
D_S	Nei's standard genetic distance
D_{SW}	Stepwise genetic distance
g	grams
IAM	Infinite allele model
ILRI	International Livestock Research Institute
Indel	Insertion or deletion
kb	kilobases
LD	Linkage disequilibrium
M	molar
MALD	Mapping by admixture-generated linkage disequilibrium
MDS	Multi-dimensional scaling

$MgCl_2$	Magnesium Chloride
mM	millimolar
mtDNA	Mitochondrial DNA
NTP	Nucleotide triphosphate
°C	Degrees centigrade
OTU	Operational taxonomic unit
PAA	Population-associated allele
PAGE	Polyacrylamide Gel Electrophoresis
PC	Principal component
PCR	Polymerase Chain Reaction
pdf	probability density function
r^2 (adj)	r-squared adjusted. Variable which takes into account the noise and number of degrees of freedom in a regression model
SFD	Scaled frequency differential
SMM	Stepwise mutation model of microsatellite evolution
SNP	Single Nucleotide Polymorphism
SSCP	Single strand conformational polymorphism
T_A	Annealing temperature
TBE	Tris borate EDTA buffer
TE	Tris EDTA buffer
UV	Ultraviolet
v.	version

SUMMARY

Genetic admixture analysis of bovine hybrid zones was carried out through implementation of a variety of statistical and analytical methods. Population genetic analysis of microsatellite variation in seven cattle breeds from a selection of locations in South Asia and in sixteen cattle populations from West Africa are presented. These are the first such studies focusing within these areas, the former of which is one of the postulated centres of cattle domestication and the latter of which contains genetically important trypanotolerant *Bos taurus* cattle breeds. Genetic relationships among the breeds were examined and levels of diversity assessed. Admixture levels were estimated using multiple approaches including the systematic selection of population associated alleles and by a model-based Bayesian admixture analysis. These analyses show a *Bos taurus* influence in the Indian sub-continent, part of a gradation which stretches from Europe through the Near East towards Indian and which may be of ancient origin. The profound effect of admixture on West African cattle populations is also revealed, and these results highlight the importance of efforts to prevent further dilution of genetically unique trypanotolerant African taurine breeds by *Bos indicus* cattle.

Closely linked X-chromosome microsatellites were genotyped in modern cattle populations with differing histories of admixture from Africa, Europe, the Near East and India. Historical hybridisation events between the two sub-species of cattle, *Bos taurus* and *Bos indicus*, have occurred in some of these populations. As linkage disequilibrium (LD) is created by admixture, cattle breeds which have been created by recent hybridisation events are expected to have higher levels of LD than those that are the result of more ancient admixture. Although the manner in which admixture occurs will affect this measure, it remains useful to estimate an effective time since admixture and could be corrected for multiple waves of introgression. This study reveals very high levels of LD among linked markers in hybrid West African cattle breeds, which are indicative of continuous gene flow into these populations and may have implications for the preservation of the trypanotolerant trait of indigenous West African breeds.

Haplotype breakdown occurs as an admixed population ages, and intact chromosomal segments from contributing parental populations are broken up by recombination. The length of intact ancestral haplotypes is likely to be critical when planning any studies using

admixture-generated linkage disequilibrium to map genes involved in disease phenotypes. Therefore closely linked microsatellites were used to examine how long haplotypes are likely to remain intact in hybrid African cattle populations by assigning ancestry in a probablistic manner to mapped microsatellites that are all located within a discrete X-chromosome region. West African hybrid populations were found to have much shorter intact chromosomal segments than more recently introgressed populations in the south of the continent. This genetic signal corresponds to the contrasting histories of the two regions. Introgression of *Bos indicus* into southern Africa has been recent, and mediated through established sedentary cultures. In contrast, genetic input to the north is older and has been facilitated by a dynamic pastoralist economy and well established, ancient trading routes. Application of this method to other areas of the genome will produce a more detailed picture of genome structure, and hence population history.

The population structure that is revealed by Bayesian inference is compared for the three datasets, revealing genuine differentiation between different cattle lineages and regional groups. Consistent trends among three large datasets were evident, with the lowest level of population structuring present on the African continent. There have been many different methods developed to estimate genetic diversity, genetic distance and admixture using microsatellite data. However, different sets of microsatellite markers will produce differing estimates for any variable. In order to combine estimates of variables obtained using different datasets, the positive correlation between those estimates obtained for populations or breeds that have been genotyped for different sets of markers was exploited by regression analysis. Mean number of alleles, unbiased heterozygosity and Bayesian admixture estimates for three cattle datasets are used as examples of how different microsatellite datasets can be successfully combined. Synthetic maps, incorporating all data points from the three datasets, were produced to depict the geographic pattern of these variables. Genetic diversity levels were highest in the Near East, and three axes of admixture were described. In general, the accuracy of this method will be improved by enlarged datasets, and could be applied to other species. In terms of the study of cattle diversity, future research will include breeds from areas such as China, Mongolia and Bhutan to produce more extensive maps of global admixture and diversity levels.

CHAPTER 1

GENERAL INTRODUCTION

1.1 INTRODUCTION

"To understand what makes us unique, both as individuals and as a species, we need to consider the genome as a mosaic of discrete segments, each with its own unique history and relatedness to different contemporary and ancestral individuals"

(PAABO 2003)

In recent years it has become clear that the genome can no longer be regarded as a homogenous entity. While physical or environmental barriers, founder effects, admixture and other historical processes may cause within-population genetic structure (BALLOUX and LUGON-MOULIN 2002), hybridisation between populations also creates structured genomes, where different chromosomal sections are derived directly from ancestral populations. In the absence of subsequent introgression, every portion of the genome of an individual from a hybrid population, created with genetic input from differentiated founders, can be traced to one or other of the contributing parental groups (CHAPMAN and THOMPSON 2002).

Tracing a portion of a genome back to an ancestral population is dependent on the ability to differentiate between genetic material that is derived from each founding population. This is also critical for estimating admixture levels in populations and individuals, and remains a challenging task. Distinct ancestral populations that have merged to produce a hybrid population are likely to have more commonality than dissimilarity in terms of genetics, and hence it is necessary to first define the differences between them. The latest human genome analyses have suggested that describing this variation in terms of haplotype

"block structure" may be incredibly useful in elucidating population history (GABRIEL *et al.* 2002; STUMPF 2002), but it is not yet clear whether this will prove to be a universal feature of mammalian genomes. While genome structure in human populations is currently being investigated (GABRIEL *et al.* 2002), little is known about genome structure in other species. The Human HapMap project is characterising patterns of linkage disequilibrium in human genome projects, and a similar study may be useful in other species (CARDON and ABECASIS 2003).

Natural populations are seldom the panmitic, closed systems that are modelled in theory. Nevertheless it is clear that deciphering genome structure originates at the population level. No interpretation of genome structure is achievable without placing a genome in the context of the population from which it has been sampled, and related populations. The relationship between the sample population and other populations is also paramount to such analysis. Extant representatives of populations that historically contributed to, or were amalgamated into a modern hybrid population are often absent and add further difficulty to such studies. An account of the relationships between extant populations, and the possible historical migrations that created them, will be critical for accurate description of modern hybrid populations.

An accurate description of the ancestry of each segment of a genome may prove extremely advantageous in ascertaining the positions of disease-causing mutations and other genes of interest (BRISCOE *et al.* 1994). The basic history of domestic cattle has been relatively clearly deciphered and, as such, they represent a suitable group on which to test methods for detecting and quantifying patterns of admixture. In the present study admixture and diversity levels in worldwide, domesticated cattle populations are focussed on and findings

are related to the history of each surveyed population. There are several regions where hybrid zones have been created by crossbreeding between genetically distinct domestic cattle lineages. These remain interesting models to study the breakdown of haplotypes and assess the effect of population history on admixture.

1.2 CATTLE

1.2.1 DOMESTICATION OF CATTLE

The word *capital* shares its origins with the French word for livestock, *cheptel*, and is testament to the early importance of these animals in the accumulation of wealth (EPSTEIN and MASON 1984; MACDONALD 2000). There are estimated to be 1.2 billion of these economically important animals worldwide (PAYNE and WILSON 1999). Domestic cattle are even-toed ungulates, which belong to the artiodactyl family *Bovidae* (PAYNE and WILSON 1999). Extant cattle types are *Bos taurus* (*B. taurus*), *Bos indicus* (*B. indicus*) and *Bos bibos*. The latter group includes Bali cattle (*B. banteng*), gaur (*B. gaurus*) and mithan (*B. frontalis*) which are found in East Asia (PAYNE and HODGES 1997).

B. taurus or taurine and *B. indicus* or zebu cattle were derived from at least two different strains of wild oxen, also called aurochs (*Bos primigenius*) (CLUTTON-BROCK 1999; LOFTUS *et al.* 1994a). Both archaeological studies and genetic analyses provide convincing evidence for separate ancestry of the two cattle lineages. The two groups are morphologically distinct. *B. indicus* animals have a distinctive muscular-fatty hump over the withers, a heavy dewlap and longer ears and legs than their taurine counterparts. They generally have a narrower body and head and are often lighter in colour than humpless cattle. Furthermore, they are physiologically better adapted to a tropical or arid environment due to their lower metabolic and water turnover rates, which must be attributed to their evolution in a warm climate (CLUTTON-BROCK 1999; EPSTEIN 1971). There are also differences between the vertebrae of the two cattle lineages (GRIGSON, 2000). Humpless taurine cattle can be divided into longhorned and shorthorned types. The longhorn type was probably domesticated first and is the group from which the humpless shorthorned type is descended. The differences between the long and shorthorned taurine

are predominantly related to reductions in size as a results of the domestication process (EPSTEIN 1971).

The primary domestication centre for humpless *B. taurus*, or taurine, cattle is believed to be the Fertile Crescent, a region of Western Asia situated between the Mediterranean and Iran (EPSTEIN and MASON 1984). This area was the probable birthplace of the Neolithic transition, when many plant and animal species were first domesticated. Present archaeological evidence suggests that cattle had been domesticated in this region by c. 8000 BP (before present). The original domesticates are thought be descended from the wild *B. primigenius primigenius*, which they resemble in several morphological characteristics (CLUTTON-BROCK 1999).

B. indicus, humped zebu or indicine cattle, are hypothesised to have later domestic origins in the Baluchistan region of Pakistan from local wild aurochs, B. *primigenius namadicus* (MEADOW 1996; ZEUNER 1963). While some archaeological evidence suggests that zebu cattle were introduced to India by immigrant pastoralists, and not domesticated independently from *B. primigenius* (PAYNE and HODGES 1997; PAYNE and WILSON 1999), genetic studies have estimated the divergence time between ancestral strains at several hundred thousand years (LOFTUS *et al.* 1994a; MACHUGH *et al.* 1997). Consequently, it seems untenable that cattle were domesticated from the same auroch sub-species, as they diverged long before domestication occurred during the Neolithic period. The data do not negate the possibility that there was some migration of taurine cattle to the Indus valley c. 5000 BP where they co-existed with locally domesticated animals for 2,000 to 3,000 years. Artefacts from the Harappan culture, which developed in the region from c. 4300 BP, show depictions of both morphological cattle types (MEADOW 1996).

1.2.2 MIGRATION OF CATTLE FROM ORIGINAL DOMESTICATION CENTRES

With the possible exception of African taurine populations, all modern cattle are considered to derive from the two main centres of domestication (**Figure 1.1**), but subsequent migrations from these two areas have changed the original distribution of domestic cattle and produced complex regional patterns of introgression, admixture and diversity (**Figure 1.2**).

1.2.2.a Migration to Europe

Taurine cattle are believed to have been introduced from the Near East to Europe by advancing pastoralists and had little or no input from local wild oxen (EPSTEIN and MASON 1984; TROY *et al.* 2001). Several different routes have been suggested, including a Bosporus route to the Danube valley and central Europe, a littoral route along the Mediterranean and direct transfer from the North African coast to Iberia via the straits of Gibraltar and to the British Isles (EDWARDS 2003; PAYNE and HODGES 1997). African mitochondrial DNA (mtDNA) has been detected in modern Portuguese cattle, which supports the theory of an African influence in Iberian cattle (CYMBRON *et al.* 1999). All other European cattle mtDNA represent a subset of the mtDNA diversity that is present in the Near East. This is not surprising as diversity is generally lost with transfer from a centre of domestication (JORDE *et al.* 1998).

1.2.2.b Migration to Africa

It has been broadly accepted that domesticated longhorned taurine cattle were introduced into Africa via Egypt from the Near East during the period that they were being established in Europe (EPSTEIN 1971; EPSTEIN and MASON 1984). From Egypt, these original introductions spread south along the Nile Valley and west into the coastal areas of North

Africa and West Africa (CLUTTON-BROCK 1999). Subsequent introduction of shorthorned taurine (c. 4500 BP) and then zebu cattle (c. 3500 BP), followed by crossbreeding between them, was believed to have produced all modern breeds (EPSTEIN and MASON 1984).

However, some commentators have speculated that domestication from local aurochs (*B. primigenius opisthonomous*) occurred independently on the African continent. There is archaeological evidence for the presence of bovids in Egypt up to 3,000 years before the appearance of domestic cattle in the Nile Valley 7000-6,500 BP (EPSTEIN and MASON 1984; WENDORF and SCHILD 1994), and for shorthorned domesticates at Capéletti in Algeria c. 5100 BP (HASSAN 2000). Some authors have speculated that the dry conditions in the eastern Sahara at the time that the first domestic bovids occurred, signal that wild aurochs could not have survived (WENDORF and SCHILD 1994). The domestication of these early cattle may have been precipitated by the need for a predictable food source under incredibly dry conditions (MARSHALL and HILDEBRAND 2002). It is possible that an initial, limited introduction of *B. taurus* animals from the Near East occurred via Egypt, but the weight of evidence now supports local domestication in North Africa (HANOTTE *et al.* 2002). It is also possible that secondary introductions of Near Eastern shorthorns also took place. African *B. taurus* cattle had reached West Africa by c. 4000 BP and these may represent the ancestors of the present West African taurine breeds which are trypanotolerant (MACDONALD and HUTTON MACDONALD 2000).

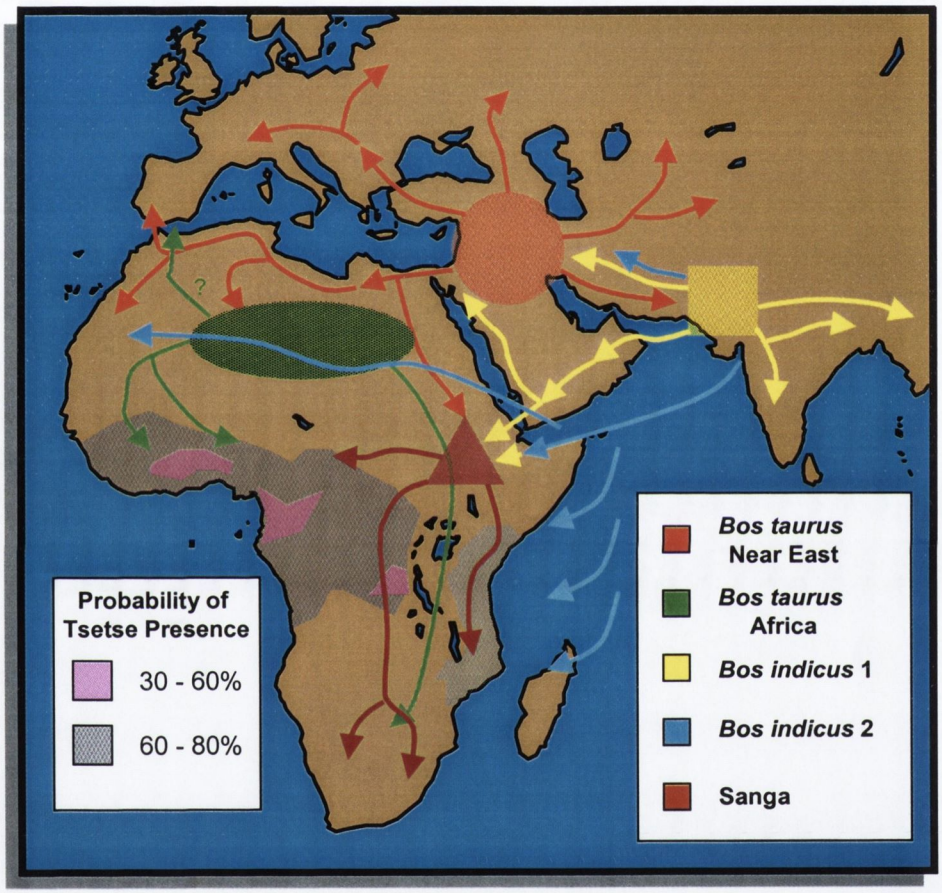
The introduction of *B. indicus* cattle into Northeast and East Africa has occurred almost continuously since Semetic-speaking peoples first brought them along the Suez from Arabia c. 2500 BP. Previously, both long and shorthorned cattle were the dominant types in the region (ROBERTSHAW 1989), but now zebu, longhorn and shorthorn have been crossed

and interbred to produce new sanga and zebu breeds. Sanga are zebu/taurine hybrids, which are thought to have been produced from the limited original crossbreeding between *B. taurus* and *B. indicus* cattle in East Africa (CLUTTON-BROCK 1989). The name is thus used to describe the predominantly taurine African breeds that are found in eastern and southern Africa (HANOTTE *et al.* 2002), and this is how it will be used in this thesis. It is likely that importation of these animals from Arabia increased as it became clear that zebu breeds thrived in the hot dry climate of the area. It is also probable that they were introduced by sea from the Indian sub-continent (HANOTTE *et al.* 2002; PAYNE and WILSON 1999).

Although the archaeological evidence remains inconclusive, there is persuasive genetic evidence for an African domestication event. MtDNA studies (BRADLEY *et al.* 1996; LOFTUS *et al.* 1994b; TROY *et al.* 2001) show consistent differences between native African and European or Near Eastern taurine cattle that likely predate the development of agriculture and domestication. All African cattle, with both zebu and taurine phenotypes, share the same group of *Bos taurus* mtDNA haplotypes. This supports the thesis that *B. indicus* input to Africa is the result of introgression into the native breeds, and not replacement (BRADLEY *et al.* 1996). Troy and colleagues (2001) found that the predominant African mtDNA is one that is absent in Europe and uncommon in the Near East. Nuclear genetic markers including Y-chromosome polymorphisms and autosomal microsatellites (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997) have been used to show that there has been a more recent widespread introgression of zebu cattle into the African continent, and that indigenous African taurine cattle only remain in certain parts of North and West Africa. The accumulated archaeological and genetic evidence indicates that indigenous African cattle are the descendants of an independent domestication event of a

local wild auroch species that was more closely related to the ancestor of European taurine cattle than that of Indian *B. indicus* cattle.

Figure 1.1. *Map showing domestication centres and subsequent routes of domestic cattle (after MacHugh et al 1996). Regions where the tsetse fly is endemic are also highlighted (FAO 2003b). The weight of current evidence supports three domestication events with subsequent mixing between them.*

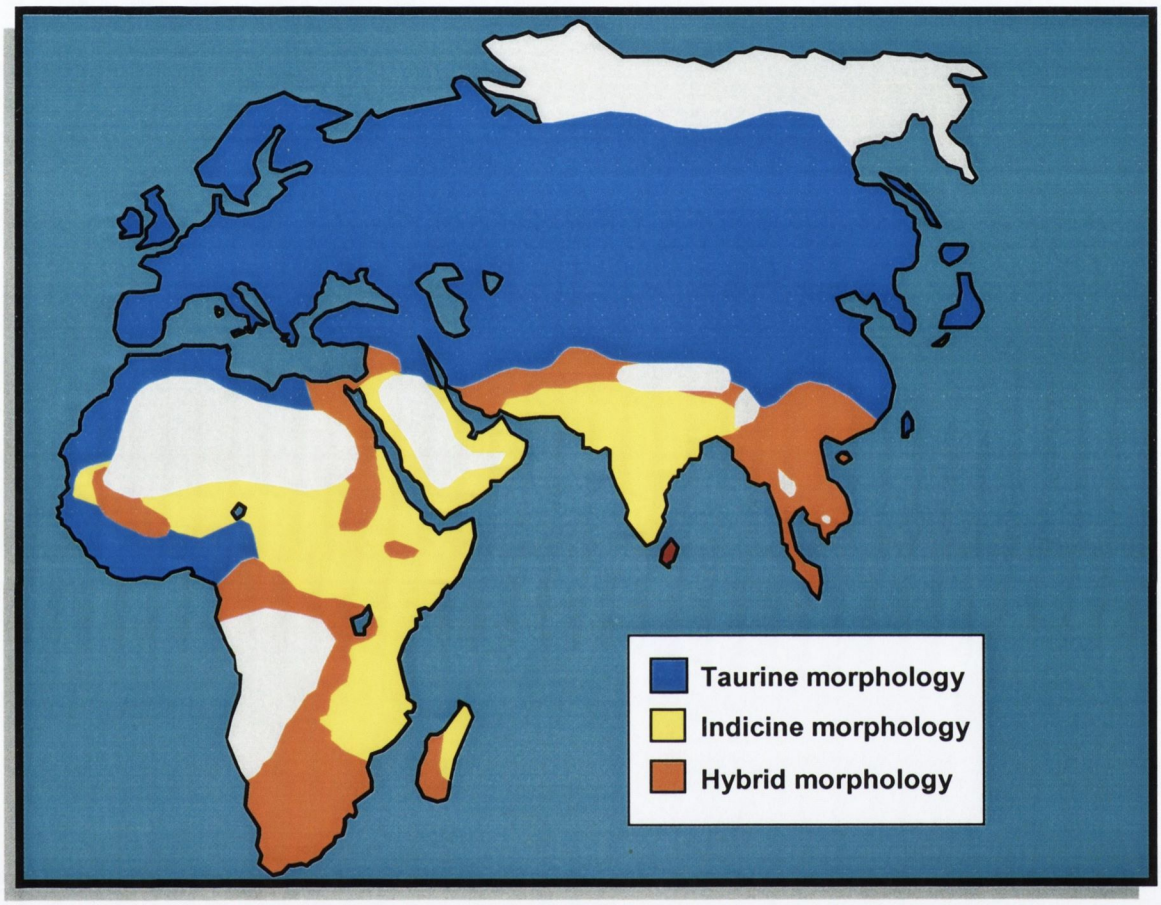


1.2.2.c Migration to India and the Near East

The *B. taurus* cattle that were originally domesticated in Western Asia and the Fertile Crescent have today undergone extensive introgression from *B. indicus* (LOFTUS *et al.* 1999). Genetic data from autosomal markers reveal a cline of zebu influence from East to West. There is some evidence for minor and possibly ancient input from Near Eastern

taurine animals into Indian *B. indicus* populations, even though these animals all have indicine phenotypes (KUMAR *et al.* 2003).

Figure 1.2 *Distribution of modern domestic cattle types (PAYNE 1970).*



There is a considerable debate about how to classify domestic descendants of a wild species. In this thesis the Linnæan taxonomy with two domesticated species, *B. taurus* (taurine) and *B. indicus* (zebu), is followed (HANOTTE *et al.* 2002).

1.3 LINKAGE DISEQUILIBRIUM

Gametic disequilibrium is defined as the non-random union of alleles in a gamete (HEDRICK 2000), but the term linkage disequilibrium (LD) is more often used to describe the non-random associations between either linked or non-linked alleles. This effect is not evenly dispersed throughout the genome and is caused by recent mutations at a locus, population founder effects, inbreeding, selection and population expansions (FREIMER *et al.* 1997; LAAN and PAABO 1997). Admixture also creates linkage disequilibrium, the physical extent of which will depend on effective population size and the local recombination rate (CHAKRABORTY and WEISS 1988; HUTTLEY and WILSON 2000). In a panmitic population and in the absence of selective effects, associations between non-syntenic alleles that are derived from a single parental population are halved every generation (BARTON and GALE 1993; WIEHE and SLATKIN 1998). On the other hand, markers that are closely linked will be co-inherited more often than expected by chance, given their frequencies in the population as a whole.

Surveys of LD in populations have found that its extent over chromosomal regions is heterogeneous and depends on both population and gene history (PFAFF *et al.* 2001). In fact, the variance in overall LD among chromosomes has been found to account for approximately two-thirds of the total LD variance in human populations (RYBICKI *et al.* 2002; SHIFMAN *et al.* 2003). This variation suggests that exploiting LD for mapping studies will require a precise description of LD at each locus, and in each population to be studied (BONNEN *et al.* 2002). Even those groups with apparently comparable histories of admixture may not share the same levels of genome structure and patterns of LD. Widely varying estimates of genome structure, as measured by LD, have been reported for related

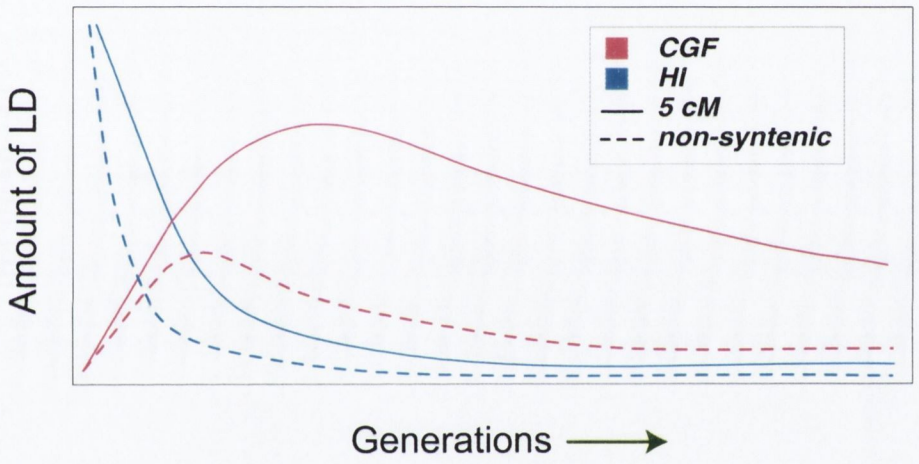
populations. Therefore both structure in the genome and admixture levels can only be assessed empirically for different populations (SHIFMAN *et al.* 2003).

Any LD that is specifically the result of admixture has been denoted ALD (admixture-generated linkage disequilibrium). This will exist between loci that genetically distinguish, by virtue of their allele frequencies, the ancestral populations (HUTTLEY *et al.* 1999; SHRIVER *et al.* 2003). ALD will be detected most successfully using polymorphic markers with a high frequency differential between the contributing parental populations (RYBICKI *et al.* 2002; SLATKIN 1994). A characteristic of newly admixed populations is that they may display higher levels of linkage disequilibrium than stable older populations (LAAN and PAABO 1997). High levels of LD have been reported in many admixed human populations (DESTRO-BISOL *et al.* 1999; LAUTENBERGER *et al.* 2000; PARRA *et al.* 2001; PRITCHARD and PRZEWORSKI 2001; WILSON and GOLDSTEIN 2000). High overall levels of LD have also been reported for livestock groups (FARNIR *et al.* 2000; MCRAE *et al.* 2002), which may be related to admixture, but which could also have been produced by other population dynamics. However, with the exception of *Homo sapiens*, background LD levels have not been assessed in many outbred mammalian species.

Different models have been proposed to explain the pattern of LD found in admixed populations (MCPEEK and STRAHS 1999; PFAFF *et al.* 2001). Decay of haplotype sharing, a measure of LD to assess the genetic distance over which the ancestral haplotype is preserved, models ALD subsequent to the introduction of a variant on an ancestral haplotype (MCPEEK and STRAHS 1999). Pfaff and colleagues put forward two models: *Hybrid Isolation* (HI) and *Continuous Gene Flow* (CGF). In the HI model the F_1 generation is the only one that receives genes directly from the two parental populations,

whereas in the CGF model there is continual gene flow from one or both of the parental populations for some time after the initial admixture event. A simulation of these two situations was carried out and the authors describe the effects of the two models on the pattern of LD (**Figure 1.3**).

Figure 1.3 (after PFAFF *et al.*, 2001). *In the hybrid isolation (HI) model (blue lines) LD begins to decrease rapidly after the initial hybridisation event, especially if the two markers are unlinked (dashed line). For two linked markers in this model (solid lines), LD will persist for longer but will decrease with every generation. In the CGF model (pink lines) the amount of LD increases initially as admixture continues and then decreases more gradually than in the HI case, especially if markers are linked).*



Pfaff and colleagues felt that the maintenance of ALD over relatively large (~10 cM) chromosomal segments is a not typical of HI but that it is a characteristic of a CGF pattern of admixture. If the frequency differential between the markers under study is high, and the recombination fraction is low this will increase the amount of LD created during the admixture process and increase the likelihood that associations between loci will be detected.

It seems clear that extensive and useful LD exists that is likely to be useful in association studies and in the elucidation of population history (LAAN and PAABO 1997; OTT 2000). Nonetheless patterns of LD are likely to show differences among populations and genomic

regions that are of practical importance in the design of such studies (STUMPF and GOLDSTEIN 2003). As such, prior assessment of LD configuration in populations of interest is recommended.

1.4 ADMIXTURE

1.4.1 A DEFINITION OF ADMIXTURE

Admixture is defined as hybridisation between interfertile but genetically distinct lineages, resulting in gene exchange to create an admixed population. This population has characteristic features, including intermediate gene frequencies between the contributing parentals (BERTORELLE and EXCOFFIER 1998). Over a short interval, however, the haplotype in a hybrid individual will consist mainly of alleles derived from one of the founding populations (MCKEIGUE 1998). Genetic ancestry may therefore vary substantially among individuals in a recently admixed population (ROSENBERG *et al.* 2002; ROSENBERG *et al.* 2001b). Admixture is also defined as the proportion of genes or gene flow from a parental to an introgressed population (HEDRICK 2000). The two models of admixture may be referred to as *Intermixture* and *Gene Flow* (LONG 1991).

1.4.2 MEASURING ADMIXTURE

1.4.2.a Methods of admixture estimation

Classical methods that were developed to measure admixture, tended to exploit the fact that the allele frequencies in hybrid populations should be linear combinations of allele frequencies in the contributing source populations (BERTORELLE and EXCOFFIER 1998). Additionally these methods assumed that the ancestries of each locus used in analysis were independent of each other, an assumption that is violated if markers are co-inherited from a single parental population. This supposition requires the admixed population to have undergone enough generations of random mating to eliminate the allelic associations that are a feature of admixture.

The extreme case of this assumption being invalidated occurs if closely linked markers are co-inherited directly on a single contiguous piece of chromosome (MCKEIGUE *et al.* 2000).

Technically this would limit the number of nuclear markers that could be included in admixture estimates to one per chromosome arm, which may not be enough to give an accurate estimate of individual admixture (REED 1973). Reed hypothesised that 18 "ideal" loci would be necessary to accurately estimate individual admixture. "Ideal" loci were modelled as diallelic loci with a frequency differential (δ) of one, or monomorphic in the two parental populations.

Traditional admixture estimators tend to equate an observed quantity directly with theoretical expectation, and may derive analytical or simulated variances around those estimates. Likelihood methods, which calculate the relative probabilities of the observed data given different parameter values, have also been employed in admixture estimation. Notation for the different measures varies between publications, however an attempt has been made to use either the original or most ubiquitous denomination. *mR* is a simple and direct least-squares linear-regression frequency based estimator (ROBERTS and HIORNS 1965), that is still commonly used (HANOTTE *et al.* 2002). The weighted-least squares (WLS) model is an extension of *mR*, which incorporates the effect of the evolutionary and sampling variance in the admixture estimates and a χ^2-test of heterogeneity of admixture estimates from the different WLS (LONG 1991).

Another popular estimator is *mC*, a maximum likelihood version of Long's 1991 method (CHAKRABORTY *et al.* 1992). To account for the contribution from both parental populations the concept of *gene identity coefficient* was developed by Chakraborty (CHAKRABORTY 1975; CHAKRABORTY 1986). This method also enabled calculation of individual admixture proportions and uses the probability that two genes drawn at random from one or more populations are identical in nature. Once again, the gene frequencies of

the hybrid population are considered as intermediate to the contributing parentals. The gene identity between one parental population and the hybrid is comprised of expressions that incorporate the probability of the gene coming from that parental population and the probability of it originating in the alternative parental population. This measure has been extended to fit a multi-locus system and multi-parental admixture model (CHAKRABORTY *et al.* 1992). *mI* is a measure developed for an island model, when admixture is the result of an instant input from an introgressing population, and is unlikely to be applicable in many situations (DI BENEDETTO *et al.* 2001).

More recent estimators have been developed that are based on the estimation of the mean coalescence times between pairs of alleles drawn from the three populations of interest (BERTORELLE and EXCOFFIER 1998). *mX* is based on mean coalescent times of markers within the admixed population, and *mY* is based on mean coalescent times for markers between the admixed and parental populations (BERTORELLE and EXCOFFIER 1998). The latter two methods permit the molecular differences between alleles to be treated under either the infinite allele or stepwise mutation models (see Section 1.7) (DI BENEDETTO *et al.* 2001), and have been extended for use with multiple parental populations (DUPANLOUP and BERTORELLE 2001).

Confusingly, several different software packages to calculate the various admixture estimates parameters have been called ADMIX. These are: ADMIX.PAS which employs Long's weighted-least squares method, ADMIX2.FOR which uses Chakraborty's gene identity method, ADMIX *v.* 1.0 calculates mR, mC and mY (www2.unife.it/genetica/Giorgio/giorgio.html), and ADMIX *v.* 2.0, which implements Bertorelle's

approach for multiple parental populations (http://www2.unife.it/genetica/Isabelle /admix2_0.html).

1.4.2.b Population-associated alleles and admixture estimation

"Private" (NEEL 1973), ideal (REED 1973), "unique" (CHAKRABORTY *et al.* 1991) or population-specific alleles (SHRIVER *et al.* 1997) have been widely used in the study of populations and individuals with disparate ancestry. These alleles are used to characterise different populations by virtue of their frequency differential. They may occur at a very high frequency in a certain population and be used to typify that population. The presence of that allele in other populations may be a "signature" of contact with the originally characterised population. When present in an admixed sample therefore, these alleles may enable the quantification of the amount of input from that parental population.

The most useful population-specific alleles are those that have the highest frequency differential (δ) between two populations. The precision of an admixture estimate is very dependent on the value of δ (CHAKRABORTY *et al.* 1991), although the polymorphism level and number of markers used are also critical (CHIKHI *et al.* 2001). Unfortunately, however, delta values must be estimated in samples of modern descendants of the founding populations that contributed to the admixed population. Often it is difficult to decide which extant populations represent the best models of the ancestral parental populations (BIERNE *et al.* 2003), and allele frequencies may have changed in the modern descendants as a result of drift, mutation, selection of other demographic events that have occurred since the original admixture event (MCKEIGUE *et al.* 2000). For multi-allelic systems, such as microsatellites, δ is calculated as the sum of like-signed individual allelic delta values.

The most informative alleles for the study of hybrid populations are those that have the largest allele frequency differences between the parental populations (GARNER and SLATKIN 2003). Ideally these alleles would be monophyletic, as homoplasy, caused by parallel mutation, reversion and convergent evolution, would mean that alleles would be identical in state but not identical by descent (ESTOUP *et al.* 1999; THOMPSON and NEEL 1997). However, if alleles are neither fixed in the parental population nor absent in the comparison populations their frequency differential may still enable their use to infer population history. They are thus referred to as ancestry informative markers or population associated alleles (PAAs) (PFAFF *et al.* 2001; SHRIVER *et al.* 2003). Nevertheless, the small frequency differentials between some microsatellite alleles may mean that they are not that useful for characterising parental populations and estimating ancestry proportions in hybrid populations.

In human populations only 10-20% of variation is between populations, hence it is difficult to find markers that fit the criteria of a population-specific allele (DEAN *et al.* 1994). In fact, 95% of arbitrarily identified bialleleic markers have $\delta < 0.5$ (DEAN *et al.* 1994). Some authors have made the recommendation, based on surveys of human populations, that markers can be successfully used, as long as δ is greater than 0.5 (MCKEIGUE 1998; SHRIVER *et al.* 1997). However, Collins-Schramm and colleagues have estimated admixture in African American and Mexican American human populations employing analogous ethnic-difference markers, with δ of >0.30 (COLLINS-SCHRAMM *et al.* 2002). It has been well established, on the other hand, that it is possible to find alleles with a high frequency differential between or that are unique to either *B. taurus* or *B. indicus* (LOFTUS *et al.* 1999; MACHUGH *et al.* 1997).

The major split in domestic cattle phylogeny is between the taurine and zebu lineages (BRADLEY *et al.* 1998; BRADLEY *et al.* 1996; LOFTUS *et al.* 1999; LOFTUS *et al.* 1994a; MACHUGH *et al.* 1997). These species are sufficiently different such that there are many species-specific microsatellite alleles. In general these have been termed zebu-diagnostic alleles (MACHUGH *et al.* 1997) and are directly comparable to the population specific alleles sought by human geneticists as mentioned above. However it is also possible to find taurine-specific alleles. In summary therefore, the accuracy of estimating admixture based on PAAs depends on correct specification of these alleles. The error in admixture estimates may be further confounded by the inaccurate selection of modern representatives of contributing parental populations and/or gene flow in those parents since the admixture event took place.

1.4.2.c Bayesian methods to estimate admixture

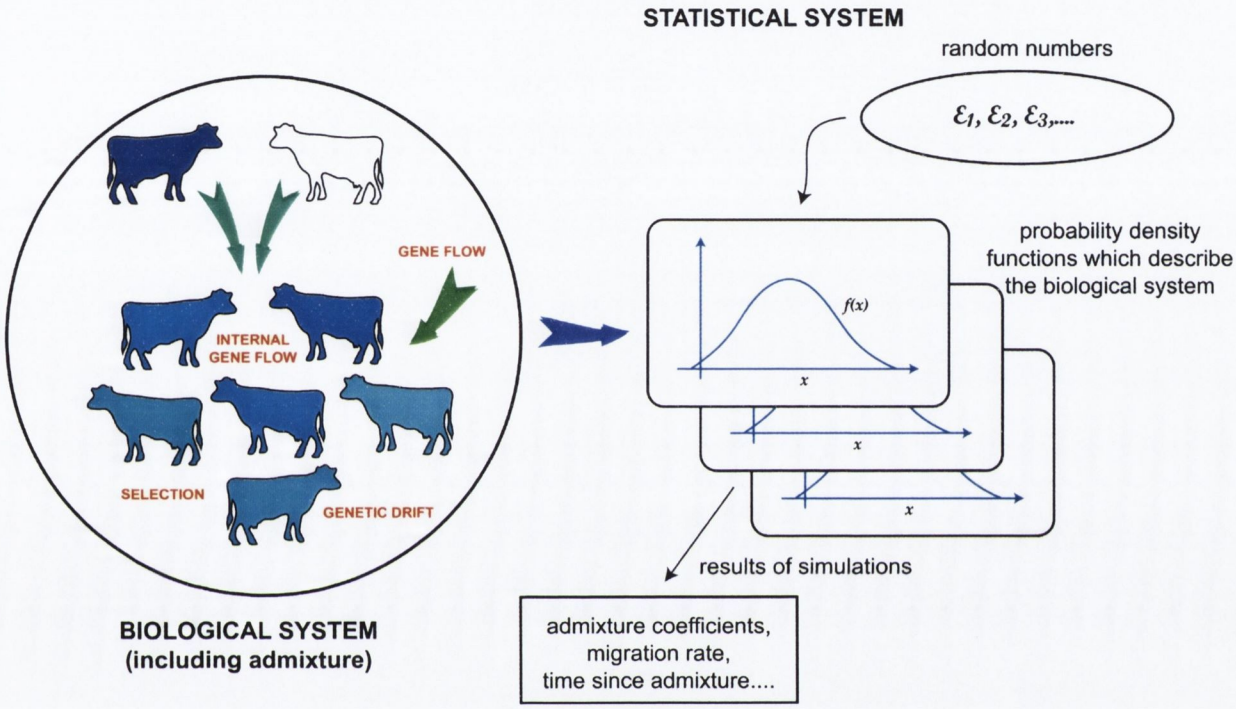
Latest methods to estimate admixture adopt a Bayesian approach, employing prior information about germane parameters to estimate admixture proportions. Bayes theorem (**Equation 1.1**) employs *priors*, which summarise previously known information about the parameter of interest (ψ), and are expressed as a *probability density function*, $p(\psi)$ (**Figure 1.4**). If no information about a particular distribution is available a flat prior may be employed (CHIKHI *et al.* 2001). The probability of the observing the data under the statistical model, $p(D|\psi)$, and the probability of the data $p(D)$ are included with the prior to estimate the *posterior* pdf, the probability distribution of ψ, given the observed data. This is calculated by:

$$p(\psi|D) = \left(\frac{p(D|\psi)p(\psi)}{p(D)} \right) \qquad \textbf{Equation 1.1}$$

If ψ represents a group of parameters, the distribution of a specific one may be estimated by averaging across all others (CHIKHI *et al.* 2001).

Models that incorporate multiple parameters such as genealogical information or genetic drift are likely to require excessive computation. However, executing these models has been made possible through the implementation of *Markov Chain Monte Carlo* (MCMC) methods. Stanislaw Ulam, John von Neumann and Nicholas Metropolis apparently coined the term "Monte Carlo" during the infamous Manhattan Project of World War II, as a Los Alamos code word for the stochastic simulations they applied to building atomic bombs (BALAZS *et al.* 2000). Their methods, involving the laws of probability, were aptly named after Monaco's renowned gaming destination and the moniker has stuck. Monte Carlo simulations are iterative stochastic algorithms that use random numbers to generate a sample population of the system, from which properties can be determined. This reduces the amount of computation required to assess all possible results. An essential component of the simulation is the ability to describe a biological process by one or more probability density functions (**Figure 1.4**). By modelling the process as such functions, which may have their origins in experimental data or in a theoretical model, it is possible to sample an outcome from them, thereby simulating the actual process (CSEP 1996). Simply, Monte Carlo allows approximation of an expectation using the sample mean of a function of simulation random variables (ANDERSON 1999).

Figure 1.4 *(modified from CSEP, 1996) Monte Carlo simulation, as applied to a biological system. Assuming that the evolution of a complex biological system can be described by probability density functions (pdfs), then the Monte Carlo simulation can proceed by sampling from these pdfs. This necessitates a fast and effective way to generate uniformly distributed random numbers. The outcomes of these random samplings, or trials, must be accumulated or tallied in an appropriate manner to produce the desired result, but the essential characteristic of Monte Carlo is the use of random sampling techniques to arrive at a solution for the biological problem (CSEP 1996).*



The random nature of Monte Carlo simulations makes them useful for sampling probability space. A continuum of estimates is built whereby known transition probabilities are used to estimate the current best estimate from the solution obtained from previous iteration. The new solution is rejected if it is not an improvement on than the previous and the latter is restored. The future state of the sequence depends only on the current state and not on the past order of states (CHAKRABORTY 2002). The sequence of best estimates is defined as a Markovian continuous general state chain and reaches convergence more rapidly that a conventional sampling method (DESAI and GLYNN 2001). Markov chains are valuable

tools for the generation of random samples from an incompletely known posterior distribution (CHAKRABORTY 2002).

Numerous methods to estimate admixture proportions using Bayesian inference have emerged in recent years. Statistical methods have been developed by McKeigue et al. (2000) to use allele frequency data to search for significant associations between marker loci and disease genes. He uses the allele frequency dependent measure *f*, the Wahlund variance, which estimates the fraction of ancestry information extracted by a bi-allelic marker when the prior probabilities of each state of ancestry are equal (MCKEIGUE *et al.* 2000). The Wahlund effect is characterised by an excess of homozygotes and a dearth of heterozygotes in a population recently created by the union of two genetically differentiated sub-samples (HEDRICK 2000). Polymorphic microsatellites can be condensed into bi-allelic systems using the methods outlined above (SHRIVER *et al.* 1997), and furthermore the algorithm has the potential to take into account some of the stochasticity in the allele frequencies of the modern representatives of the parental populations (CHIKHI *et al.* 2001; MCKEIGUE *et al.* 2000). Software to implement this method (ADMIXMAP) is available at http://www.lshtm.ac.uk/eph/eu/GeneticEpidemiologyGroup.htm.

A genealogical likelihood-based approach to estimating admixture has been developed by Chikhi and colleagues (2001). The algorithm takes into account drift in a hybrid population and two parental populations, and estimates the time since admixture occurred, conditioned on population size. Coalescent theory is used to simulate genealogy trees based on different demographic models. The Bayesian method includes alleles frequencies

and population sizes as priors. Software to implement this method (LEA) is available at *http://www.rubic.rdg.ac.uk/~mab/* (CHIKHI *et al.* 2001).

STRUCTURE (PRITCHARD *et al.* 2000a) employs Bayes theorem to detect population structure, assign individuals to populations in a probabilistic manner, and detect migrants and admixed individuals. The algorithm is not conditioned on a predefined mutational process and can be used with most genetic markers. The most unique feature of this method is that parental populations are not predefined. Both parental and hybrid populations are inferred, and are based on a characteristic set of allele frequencies. The model infers different populations by assuming that non-admixed populations are in Hardy-Weinberg equilibrium, and thereby accounts for linkage or Hardy-Weinberg disequilibrium by the introduction of population structure. Discrete groups that have limited disequilibrium are therefore detected (PRITCHARD *et al.* 2000a). This method has been extended to include linked markers and can simultaneously estimate recombination rate (FALUSH *et al.* 2003). Software is available at *http://pritch.bsd.uchicago.edu*.

While methods to estimate admixture have become increasingly sophisticated, none have managed to incorporate all of the desirable variables. As represented in **Figure 1.4**, processes that occur after the admixture event may confound admixture estimates. All of the models discussed above assume limited gene flow between the parental populations into the hybrid population since the significant contact, and do not cater for a CGF model (Section 1.3). Except for the STRUCTURE method, all other estimators require specification of parental populations, which need to be carefully determined to ensure accuracy.

1.4.2.d Other methods of admixture estimation

As a hybrid individual will have alleles more common with one ancestral population over any given short genome interval, linkage disequilibrium methods can be used to estimate admixture levels (BRISCOE *et al.* 1994; CHAKRABORTY and WEISS 1988; KAPLAN *et al.* 1998). Deviations from Hardy-Weinberg equilibrium (HWE) have been also used to test for population admixture using the coefficient D or normalised versions of D (SLATKIN 1994; ZAPATA 2000). However, Deng and colleagues found that the power of this test to detect admixture is usually small and that a LD method for detecting population admixture is more robust and yields fewer false positives (DENG *et al.* 2001).

Estoup and colleagues (1999) have suggested using juxtaposed microsatellite systems to circumvent some of the problems with mis-assignment of diagnostic alleles and parallel mutation in admixture studies. They suggest the use of two polymorphic loci within 200kb of each other to assign ancestry of a particular chromosomal region to a parental population. This may be particularly useful when working with diploid systems as if three of the alleles are from one progenitor group and the other is not distinguishable, then the region can be confidently assigned that that group. As the number of bovine markers available increases this may be a system that can be used in cattle admixture studies. SNPs may have disequilibria at large distances which could produce associations between markers that are not physically linked, so it may be optimal to use them in conjunction with microsatellite systems. It may possible to sequence the flanking regions of bovine microsatellites to locate population-specific SNPs that could be used as a 2-marker system with the microsatellite to add confidence to the ancestral designation of that locus (COLLINS *et al.* 1999).

1.4.3 WHY THE INTEREST IN ADMIXED POPULATIONS?

Interest in admixed populations has increased because of their potential use in mapping genes for complex traits such as hypertension and non-insulin dependent diabetes that occur at high frequencies in certain populations (KAPLAN *et al.* 1998; MCKEIGUE 1997). Previously, families with admixed ancestry were excluded from traditional linkage analysis (BARNHOLTZ-SLOAN *et al.* 2001). However, in recent years it has become clear that populations with clearly defined histories of admixture may prove incredibly useful in gene mapping studies, as a result of their higher that usual levels of LD (LAAN and PAABO 1998). If admixture has occurred recently, and the levels of LD are high, this will greatly reduce the number of markers required for a genome-wide linkage scan for a gene affecting the trait of interest (MCKEIGUE 1997; MCKEIGUE 1998).

1.5 MAPPING BY ADMIXTURE GENERATED LINKAGE DISEQUILIBRIUM

One of the primary reasons to define and understand the underlying block structure in the genome of any species is the potential to use this information for the mapping of genes involved in both simple and complex disorders (BRISCOE *et al.* 1994; CHAKRABORTY and WEISS 1988). This method involves comparing the frequencies of neutral markers between affected and unaffected individuals in a hybrid population that was created with input from two parental populations with differing frequencies of certain genetic diseases or hereditary traits. Disease alleles with a recent common ancestor will be surrounded by a larger ancestral haplotype, as will disease alleles in a newly admixed population (NORDBORG and TAVARE 2002).

Figure 1.5 shows that individuals with a larger proportion of their genome derived from one parental population have a higher chance of being affected by the traits that occur at a high frequency in that ancestral population (SHRIVER *et al.* 2003). If an allele at a neutral marker is physically linked to the trait-causing variant, this allele will occur at a high frequency in the affected parental population. A higher than expected frequency of this allele in the affected members of the hybrid population compared to the unaffected members, given their admixture coefficients, indicates that the neutral marker and the disease-causing gene may have been co-inherited and may be closely physically linked. However, admixture may also create linkage disequilibrium between markers that are not physically linked, but have both been inherited on separate chromosome segments. These false positive associations may confound the discovery of genuine associations between neutral markers and the gene of interest (BARNHOLTZ-SLOAN *et al.* 2001; MCKEIGUE 1997; MCKEIGUE 1998; PRITCHARD *et al.* 2000b; ROSENBERG *et al.* 2002).

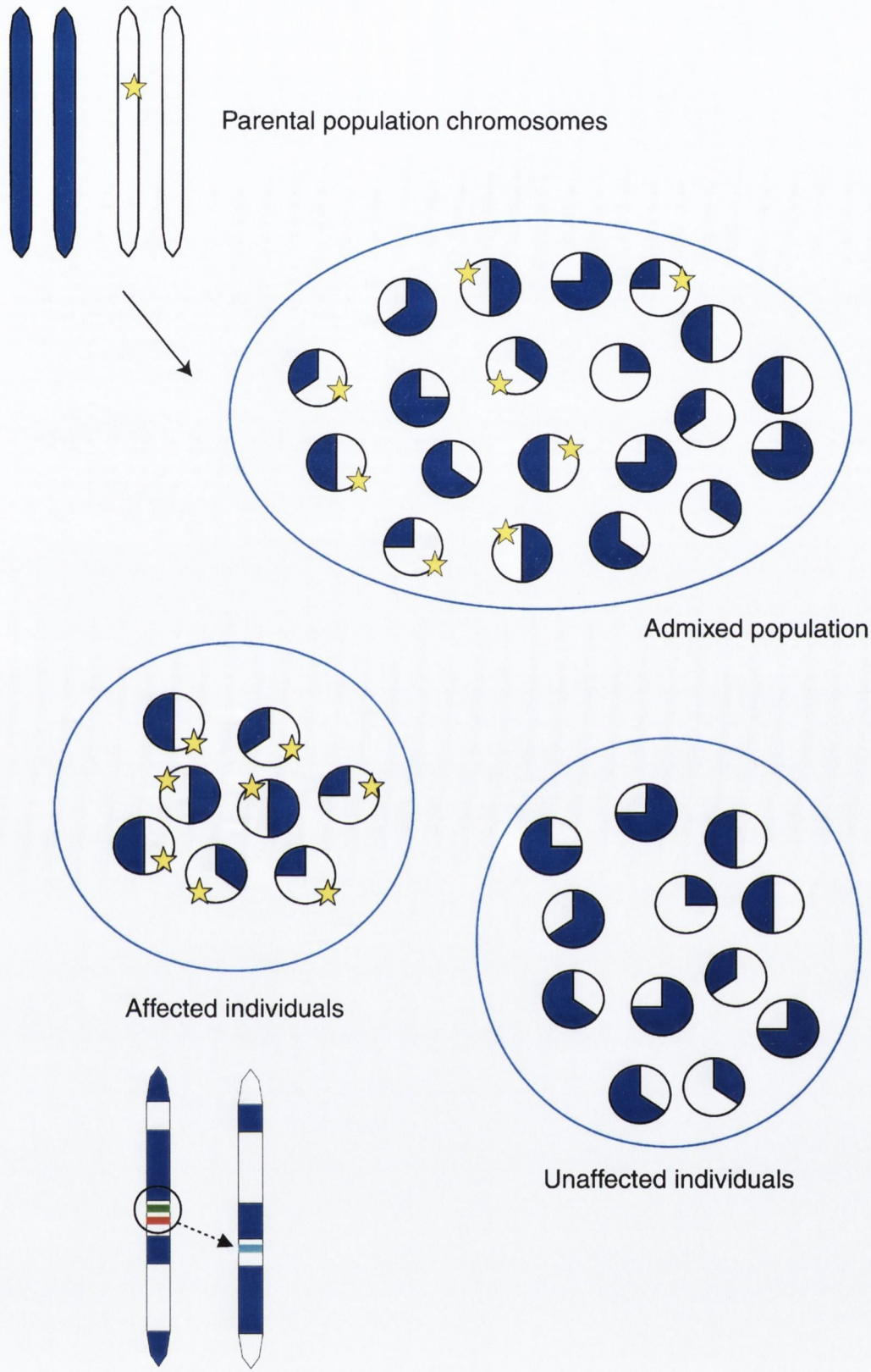


Figure 1.5. *Schematic illustrating the principles of mapping by admixture-generated linkage disequilibrium. The top left-hand corner show chromosomes from the genetically distinct populations that contributed to the admixed population shown in the first ellipse A hereditary disease, D, is present at a high frequency in the*

parental population that is shown in white and caused by a mutation marked with a yellow star. Each individual in the admixed population is represented by a pie chart. The pie chart colour divisions show the proportion of each individual's genome that is derived from each of the two parental populations. Individuals that are affected by the condition D are shown marked with a yellow star. The admixed population may be divided into affected and unaffected individuals. In general individuals with a greater proportion of their genome from the affected parent are more likely to have inherited condition D. A random genome scan of all individuals in the admixed population may reveal that there are alleles that occur at a high frequency in the affected parent and a higher frequency than would be expected by chance in the affected admixed individuals, compared to the unaffected admixed individuals. The disease-causing variant may be physically linked to one or more of these markers as they have been co-inherited on one segment of chromosome. This is shown by the left-hand chromosome in the lower part of the figure. The green and red marks represent the neutral marker and the disease-causing variant respectively. False positive scores may be obtained if a neutral marker on another chromosome in the same individual (shown by the right-hand chromosome in the lower part of the figure), has a significant association with the disease-causing mutation because they have been co-inherited on different chromosome blocks from the affected parent.

The success of this method of gene mapping is highly dependant on the accurate estimation of individual admixture levels, population admixture levels and allele frequencies for both neutral markers and trait-causing genes. In human populations it has been noted that employing long range LD, created in recently admixed populations, and markers with high δ values is likely to be profitable for mapping studies. Markers lacking with low values of δ have lower correlations for both recombination and physical distance (RYBICKI *et al.* 2002). While there is a relative paucity of suitable markers in human populations, in cattle they abound, especially between *B. taurus/B. indicus* hybrids. No previous study of LD in cattle has included these diverse populations. Therefore this survey represents the first assessment of LD of cattle from outside Europe and includes the threatened trypanotolerant breeds of West Africa and both isolated and outbred populations, a combination of which is likely to be advantageous for gene mapping (CHAPMAN and THOMPSON 2001).

1.6 CATTLE AND TRYPANOSOMIASIS

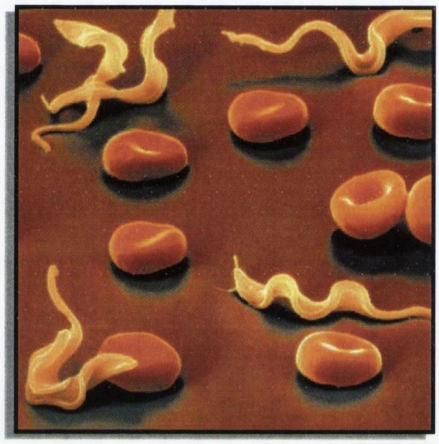
Trypanosomiasis is spread by the tsetse fly (*Glossina* sp., **Figure 1.6**) in Sub-Saharan Africa, where it represents a major constraint to socio-economic development (KABAYO 2002). The flies represent the vector for the transmission of this potentially fatal wasting disease. The areas where tsetse flies are endemic are shown in the grey and pink areas of **Figure 1.1**. The disease is also known as sleeping sickness and nagana (in cattle) and affects both livestock and humans (LEGROS *et al.* 2002; STICH *et al.* 2002).

Figure 1.6 *Tsetse fly on human skin. Tsetse flies are the vector for the spread of the trypanosome parasites, which enter the blood stream and cause the wasting disease, trypanosomiasis.*

The impact of animal trypanosomiasis on agriculture in sub-Saharan Africa is not inconsequential. Food production may be reduced by as much as 50% as a result of 3 million livestock deaths each year, a 20% loss in calving, a 25% reduction in milk yields and an overall 50% reduction in herd numbers (FAO 2003b). In East Africa trypanosomiasis is manifest as an acute and fatal disease, whereas the West African form tends to more chronic and debilitating, causing infected animals to be lost for milk and meat production. The disease is caused by extracellular parasites, predominantly *Trypanosoma congolense, T. brucei* and *T. vivax* (**Figure 1.7**). While *B. indicus* cattle are

better adapted to the increasingly dry areas of West Africa than *B. taurus* breeds, they are susceptible to trypanosomiasis and require trypanocidal drugs to survive in some areas of West Africa. Emerging resistance to these drugs in both cattle and humans represents a genuine threat to continuing successful production in the area (EISLER *et al.* 2001; MATOVU *et al.* 2001).

Figure 1.7 *Trypanosomes (T. brucei) and red blood corpuscles (from NATURE, **405**, 8th June 2000)*



Human African trypanosomiasis is one of the most important but equally most neglected tropical infections, which threatens more than 60 million people in 36 countries of sub-Saharan Africa (WHO 2002). After being brought under control in the 1960s, HAT has now re-emerged as a epidemic of immense proportions in many historic foci across Africa. With 45,000 cases officially reported in 1999 the disease has, until recently, received little attention from the international community. However, according to the World Health Organization, about 500 000 people already carry trypanosomes and will die if left untreated. (FAO 2003b; STICH *et al.* 2002).

Novel strategies to combat the disease are coordinated by the Pan African Tsetse and Trypanosomiasis Eradication Campaign, which has been developed by African governments to eradicate the flies from areas of tsetse infestation (FAO 2003b).

Controversial sterile insect techniques, which have been questioned on historical, ecological, logistical and financial grounds, may be employed to ultimately create tsetse-free zones (KABAYO 2002; ROGERS and RANDOLPH 2002).

In the light of the difficulties that may be involved in any tsetse eradication program, the indigenous trypanotolerant cattle breeds of West Africa represent an especially important genetic resource that could be used to discover the genetic background to trypanotolerance. Discovery of the genes that provide some cattle with trypanotolerance could lead to strategies that would enable African farmers to breed cattle that are productive, and maintain the disease resistance trait. However, possibly more significantly this understanding may contribute to successfully combating the disease in humans (KEMP 2002).

1.7 MICROSATELLITES

1.7.1 INTRODUCTION TO MICROSATELLITES

Microsatellites are simple tandem repeats, usually 1-5 nucleotides long, which are found randomly dispersed throughout the nuclear genome (SCHLOTTERER 2000). These markers have proved to be eminently suitable for population genetics (GOLDSTEIN and SCHLOTTERER 1999), partially because of their rapid rate of mutation; estimated to be from 10^{-2} to 10^{-5} per locus per generation (EDWARDS *et al.* 1992; ELLEGREN 1994). This feature of microsatellites evolution makes them especially useful for elucidating relationships between closely related populations (TAKEZAKI and NEI 1996). The mutation rate is affected by both repeat length and base composition, dinucleotide repeats mutate faster than trinucleotide repeats, and sequences with a high AT content mutate faster than those with a GC content (SCHLOTTERER and TAUTZ 1992). Additionally, a positive correlation has been found between repeat number and mutation rate (BRINKMANN *et al.* 1998).

1.7.2 MICROSATELLITE MUTATION

It has been proposed that microsatellites mutate by two main mechanisms, which may act in concert. These are strand-slippage replication and unequal crossing over in meiosis (LEVINSON and GUTMAN 1987). Strand-slippage probably occurs during lagging strand synthesis when one DNA strand is static, and involves the slippage of the new DNA strand upon dissociation of a polymerase complex. This slippage creates a bulge that, upon DNA repair, may either remove or add a repeat unit (SCHLOTTERER and TAUTZ 1992). Microsatellites can present some difficulties in analysis thus different models have been proposed to account for their evolution (HEDRICK 1999; WILSON and BALDING 1998).

1.7.2.a Infinite allele model (IAM)

The infinite allele model (IAM) was developed to understand the mutational processes that occurred for allozymes (KIMURA and CROW 1964). For these markers, it was assumed that most new mutations gave rise to new alleles that would be electrophoretically distinguishable. All identical alleles are considered to be identical by descent (BALLOUX and LUGON-MOULIN 2002). While the IAM has proved very successful in explaining the observed allozyme variation (NEI 1987), it also seems to fit some microsatellite mutation simulations; especially dinucleotide repeats (SHRIVER *et al.* 1993).

1.7.2.b Stepwise mutation model (SMM)

The stepwise mutation model (SMM) (KIMURA and OHTA 1978) has been widely used for the analysis of microsatellite data. Allele length is treated as an incremental character and microsatellites are presumed to evolve by the gain or loss of single repeat units (GOLDSTEIN *et al.* 1995a; SHRIVER *et al.* 1995; SHRIVER *et al.* 1993). This model allows for homoplasy, as alleles can be identical in state, but not identical by descent (BALLOUX and LUGON-MOULIN 2002). Homoplasy leads to the underestimation of the total amount of variation and genetic distance. The mode of mutation of microsatellites suggests a bias towards an increase in repeat length (RUBINSZTEIN *et al.* 1995), but it is clear from empirical data that there is a size limitation on the number of repeats (BOWCOCK *et al.* 1994). This means that more homoplasy than predicted by the SMM is likely to occur in reality, especially as the average size of microsatellite alleles increases.

1.7.2.c Two phase model

A more recent model encompasses the mutational process of the SMM, but permits mutations of more that one repeat unit to occur. This two-phase model is based on coalescence theory and predicts the expected variance in repeat number under different

mutational processes and demographic histories (DI RIENZO *et al.* 1994). The probabilities of any mutation of being a single or multistep are p and 1-p respectively (BALLOUX and LUGON-MOULIN 2002). The data of Di Rienzo et al. (1994) also suggest that microsatellite loci do not all evolve at the identical rates. They observed that an increase in the variance of repeat number led to an increase in the frequency and/or size of multi-step mutations. This result is congruent with the multi-step mutations being caused by unequal crossing over. As the variance in repeat number increases, the probability of a crossing over event causing an unequal exchange greater than one repeat unit, should also increase. Moreover, variance in allele size has been found to be related to locus mutation rate (VALDES *et al.* 1993).

The evolution of microsatellites is a complex process and different mutation models may be more appropriate for different loci. Some authors have suggested that microsatellites with repeat units of different sizes should be treated as a heterogeneous group (SHRIVER *et al.* 1993). Understanding the mutational process is essential before relationships between observed variation and genetic distance or population substructure can be inferred. Imperfections in the repeat array tends to stabilize microsatellites (GOLDSTEIN and CLARK 1995), therefore composite microsatellites (core repeats that are interrupted by point mutations) will have a more IAM-like evolution process (ESTOUP *et al.* 1995a). These markers will display lower levels of homoplasy, as their complexity will decrease the likelihood of single step mutations, and may be more suitable for discerning relationships between distant groups (ESTOUP *et al.* 1995b)

1.8 OVERVIEW

In the present study I have undertaken to accurately estimate both population and individual level admixture in world cattle populations (**Chapters 2, 3 and 5**) using a variety of methods, including examination of levels of linkage disequilibrium in a mapped area of one chromosome (**Chapter 4**). Populations with different histories of admixture have been examined to ascertain whether it is possible to detect differences in chromosome block structure, with particular focus on the trypanotolerant breeds of West Africa.

CHAPTER 2

Admixture analysis of South Asian Cattle

2.1 INTRODUCTION

As explained in **Chapter 1**, both genetic and archaeological evidence supports at least two domestications of cattle from different wild progenitor aurochs races. *Bos indicus (B. indicus)*, or zebu, are postulated to have domestic origins in the Bulachistan region while *Bos taurus (B. taurus)*, also termed taurine cattle, arose in the Near East and possibly North Africa. Genetic studies have estimated the divergence time between ancestral strains at several hundred thousand years (LOFTUS *et al.* 1994a; LOFTUS *et al.* 1994b; MACHUGH *et al.* 1997).

While it is clear the two lineages are genetically distinct, archaeological evidence is inconclusive as to whether the first cattle in the Indian sub-continent were *B. taurus* (humpless) or *B. indicus* (humped). Nevertheless, it is possible that taurine cattle were initially brought from the Near East and then *B. indicus* cattle were domesticated from local aurochs in copycat operations c. 5000 BP. Depictions of the animals on excavated materials dated to that era shows phenotypes that are typical of both modern central Indian breeds, and zebu/taurine crossbreds. Introduced *B. taurus* herds may have been improved by substantial input of indigenous animals (MEADOW 1996).

However, as humpless cattle gradually spread southeast towards the Indus valley, it is likely that fertility rates decreased and mortality increased, the typical response of taurine type cattle that are introduced to a hot and humid environment. These circumstances would have meant that the numbers of *B. taurus* cattle declined, while the *B. indicus* individuals thrived successfully. In fact, there is no archaeological evidence for the presence of taurine cattle on the Indian sub-continent after around 2800 BP. Extinction of *B. taurus* almost certainly occurred before the cattle were widely distributed, but does not

preclude the possibility of remnant taurine genetic input into modern Indian cattle (PAYNE and HODGES 1997).

Genetic evidence lends credence to the hypothesis that there has been contact between the two cattle lineages since they were separately domesticated, which may be of ancient or modern origin. *B. taurus* mitochondrial DNA (mtDNA) control regions may be divided into 5 main geographically distributed haplotypes. The haplogroup that has been denominated T3 predominates in Europe and is one of the 3 haplogroups (T, T2 and T3) that are found in the Near East. T1 is almost exclusive to Africa and T4 is found only in Japan (MANNEN *et al.* 1998; TROY *et al.* 2001). MtDNA studies undertaken on the cattle of the Indo-Pakistan region have revealed the presence of taurine mitochondria in some *B. indicus* breeds: Red Sindhi, Tharparker and Hariana. A few individuals from within these breeds were found to contain the T or T2 mtDNA control region haplotype. This demonstrates the possibility of introgression of some taurine material from the Near East, but probably not from Europe (MAGEE 2002). As mtDNA is maternally inherited it only provides us with one aspect of the history of these populations. The analysis of autosomal markers in the current study should provide us with results that will complement previous mtDNA work and which can further clarify the history of these populations.

To re-examine the origins and subsequent history of zebu cattle, microsatellite allele frequency data from 20 markers that have been screened in seven breeds with origins in a variety of locations in South Asia are analysed. These data represent the first published molecular survey of cattle within the region and support earlier findings which suggest that zebu were domesticated from a separate strain of aurochs to that which gave rise to *B. taurus* (taurine) cattle (LOFTUS *et al.* 1994a). The nature of any secondary contact between

zebu cattle and other domestic strains is of interest from a historical perspective. It also has important implications for any admixture studies, where Indian cattle are often used as representative of putative parental populations (HANOTTE *et al.* 2002; LOFTUS *et al.* 1999).

The *B. indicus* and *B. taurus* genomes are sufficiently different such that there are many both population specific and population associated microsatellite alleles. Those at high frequencies in indicine populations have been termed zebu-diagnostic alleles (MACHUGH *et al.* 1997) and have been used previously, after identification by eye, to quantify the amount of zebu introgression into native taurine populations in Africa and the Near East. Mitochondrial DNA, Y-chromosome and other allele frequency based approaches have been used to support admixture estimation by zebu-specific alleles. In these studies European and African taurine and Indian zebu cattle have been used as parental populations with which to infer the amount of admixture in hybrid breeds (BRADLEY *et al.* 1994; HANOTTE *et al.* 2002; LOFTUS *et al.* 1999; MACHUGH *et al.* 1997). However, the accuracy of these methods depends on the availability and correct selection of parental populations. There is a reasonable hypothesis, given its proximity to the Eastern region of the Fertile Crescent, that *B. taurus* admixture (possibly very early in origin) may have encroached into India. However, without a description of this there is no objective method with which to choose a parental (or introgressed) *B. indicus* population without using circular argument.

New microsatellite data has been supplemented with published data and an unbiased system to select taurine-specific alleles has been devised. The frequencies of these alleles in seven Indian breeds have been examined to ascertain the degree, if any, of taurine introgression. The authenticity of our allele choice and admixture estimates has been

tested. Additionally, average genetic distance between each Indian breed and 11 European populations were examined as an alternative indication of *B. taurus* influence. Finally, Bayesian methods developed by Pritchard and colleagues (2000) were employed to estimate the posterior distributions of individual admixture. This technique, which does not require parental population identification, has been employed in similar studies in humans (ROSENBERG *et al.* 2002), wildcats (BEAUMONT *et al.* 2001) and brown trout (HANSEN *et al.* 2001), and here illustrates taurine influence in Indian *B. indicus* cattle. Traditional allele frequency based admixture measures that require pre-selection of parental populations are avoided, along with the inherent bias that they bring to admixture analysis.

2.2 MATERIALS AND METHODS

2.2.1 SAMPLES

To augment samples collected for other studies (EDWARDS *et al.* 2000; LOFTUS *et al.* 1999; RITZ *et al.* 2000) hair samples were collected from cattle breeds in the South Asia region (**Table 2.1; Figure 2.1**) ensuring, when possible that the animals selected were unrelated and considered typical of the breed. Sampling locations are shown in **Figure 2.1**. Data from the Ongole and Nellore breeds have been represented previously (LOFTUS *et al.* 1999). Information about the surveyed breeds is detailed below.

2.2.1.a Hariana

Hariana cattle are medium-sized white or grey cattle that are delicately built, with long, narrow heads. Bullocks are used for draught, while females are moderately successfully milked. They are bred throughout India, but originate from the Hariana tract of Haryana state, India (PAYNE and HODGES 1997).

2.2.1.b Sahiwal

Sahiwal cattle are large heavily built animals with short sturdy legs. They vary in colour but are often a reddish-dun colour. They have short thick horns and the massive hump of the male often falls to one side. Sahiwal cattle are named after the area from where they originate, a district of the Punjab province of Pakistan, where they are currently bred on large farms. Producing excellent milk yields, they have been exported to many countries worldwide and a National Sahiwal Stud was established in Naivasha, Kenya, in 1962 (PAYNE and HODGES 1997).

2.2.1 c Red Sindhi

Red Sindhi cows are recognised for their superior milk production among Indian breeds and have been used for upgrading of other local cattle. The animals tend to be stocky-framed and medium to small in size. The name is derived from their red coats, which are short and soft, on loose, pigmented skin. They have upward curving horns and males have a well-developed hump. The Red Sindhi breed is believed to have been first developed in Pakistan in the Karachi region. The first named herds were probably derived from the Las Bela breed from the hills of Baluchistan (PAYNE and HODGES 1997).

2.2.1.d Tharparker

Tharparker cattle are known for their high endurance levels and are a dual-purpose draught/dairy breed. They are well-built, medium sized animals with moderate sized humps. Their coats are usually grey with a light grey strip along the top line. The breed originates in the Sind desert of Pakistan and has been interbred with cattle from surrounding areas (PAYNE and HODGES 1997).

2.2.1.e Desi

Desi is a Hindi term used for small, nondescript local cattle, which tend to be herded by nomadic people, with little or no selective breeding. This term is used here for a village sample from a localised area in central India (**Figure 2.1**) (FELIUS 1995).

2.2.1.f Nellore

The Nellore breed was sampled in Brazil and is derived from Ongole cattle imported from India from 1895 onwards (FAO 2003a; FELIUS 1995).

2.2.1.g Ongole

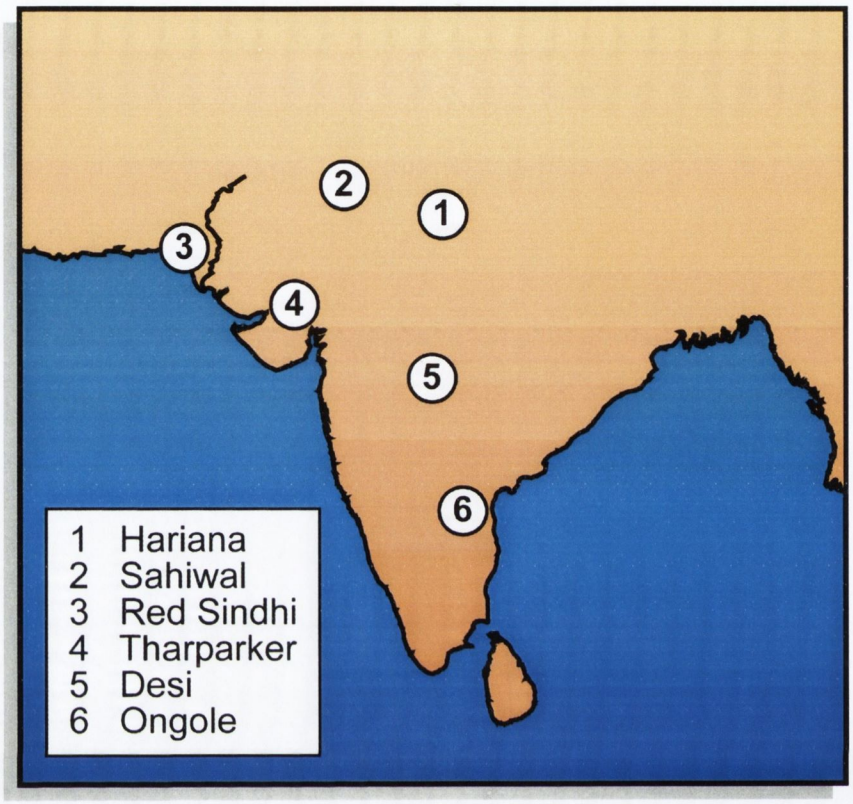
Ongole cattle are one of the largest of the Indian breeds and have been exported worldwide. Except for mature males, which may be grey, they usually have white coats. Short stumpy horns and slightly drooping ears are situated on long heads, while the dewlap tends to be large and fleshy.

2.2.2 GENOTYPING

DNA from hair was extracted and samples were screened with 20 microsatellite loci by Pushpendra Kumar (Animal Genetics Dept., Indian Veterinary Research Institute, Izatnagar, Bareilly (U.P.) India), using the procedures outlined elsewhere (LOFTUS *et al.* 1999).

The loci were selected from a list of 30 recommended for collaborative study (http://www.ri.bbsrc.ac.uk/cdiv_www/homepage.htm) and were: CSRM60, CSSM66, ETH3, ETH10, ETH152, ETH225, HEL1, HEL5, HEL9, HEL13, BMS1818, BMS1824, BMS2113, ILST005, ILST006, INRA005, INRA023, INRA032, INRA035, and INRA063. The European (*Bos taurus*) breeds that were used are: Jersey (JE), Simmental (SI), Holstein (HL), Swiss Brown (OB), Eringer (ER), Evolenard (EV), Charolais (CH), Vosges (VG), Pustertaler (PS), Pinzgauer (PG), Hungarian Grey (HG). The Near Eastern breeds that were used are: Turkish Grey (TG), Anatolian Black (AB), South Anatolian Red (SAR), East Anatolian Red (EAR), Damascus (DM), Egyptian (EG), Kurdi (KUR) and Iraqi (IR).

Figure 2.1 *Map of the Indo-Pakistan region showing the site of origin of six Indian cattle breeds. The Nellore breed was sampled in Brazil but is derived from the south Indian Ongole breed. Desi are local non-descript cattle that were sampled at this location and do not represent an official breed.*



2.2.3 ANALYSIS OF MICROSATELLITE DATA

Allele frequencies for all loci were determined directly by counting and are shown in **Appendices A1 and A2**.

2.2.3.a Mean Number of Alleles ($\overline{A}$)

Mean number of alleles per locus ($\overline{A}$) was also calculated as an index of sample diversity. $\overline{A}$ is a traditional measure of diversity that is widely used in analysis of microsatellite genotypes. It is a measure that is highly sensitive to admixture and can be useful in identifying recently admixed populations. Typically isolated breeds and breeds undergoing intense selection will have low $\overline{A}$ values while breeds that have been recently

introgressed or which are experiencing high levels of gene flow will have high $\overline{A}$ values. $\overline{A}$ is given by:

$$\overline{A} = \sum_{j-1}^{r} Aj / r \qquad \textbf{Equation 2.1}$$

where Aj is the number of distinct alleles at locus j and r is the number of loci. A size-adjusted mean number of alleles estimate was carried out using 1000 replicates of a sampling with replacement method. This was implemented using the MICROSATELLITE TOOLKIT (Park, 2001, http://oscar.gen.tcd.ie/~sdepark/ms-toolkit/index.html).

2.2.3.b Gene Diversity

Nei's unbiased gene diversity (NEI 1987) or expected heterozygosity is the probability that an individual will be heterozygous at a given locus ($\hat{h}$) or over the assayed loci ($\hat{H}$) for a multi-locus system. This genetic diversity indicator is less sensitive to sample size than the number of distinct alleles a particular locus. In an isolated population the measure will decrease with drift, as the values of p and q in Hardy–Weinberg tend towards zero or one. Population bottlenecks, inbreeding and selective breeding may also act to decrease heterozygosity, as alleles are lost from a population. Founder effects may produce populations with low heterozygosities. Any event that results in gene flow between populations, such as hybridisation and introgression events (especially continuous), will be inclined to increase the number of heterozygotes in a population.

Nei's standard unbiased heterozygosity with associated standard errors was computed for all locus/population combinations according to Nei (1987), using the MICROSATELLITE TOOLKIT (PARK 2001). An unbiased estimates of gene diversity is given by

$$\hat{h} = 2n(1 - \sum \hat{x}_i^2)/(2n-1) \qquad \textbf{Equation 2.2}$$

where

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

where $\hat{X}_{ij}$ is the frequency of genotype of A_iA_j in the sample and n is the number of individuals samples. Average gene diversity (across *r* loci) is given by:

$$\hat{H} = \sum_{j=1}^{r} \hat{h}_j / r \qquad \textbf{Equation 2.3}$$

2.2.3.c Population-Associated Alleles

B. taurus associated alleles were selected by comparing pooled European breed data with pooled data from all Indian samples. A scaled frequency differential (SFD) was calculated for each allele by the formula

$$\frac{|P_E - P_I|}{\max(P_E, P_I)} \qquad \textbf{Equation 2.4}$$

where $\boldsymbol{P_E}$ is the average frequency of the allele in the eleven European breeds and $\boldsymbol{P_I}$ is the average frequency in the seven Indian breeds assayed. The denominator denotes the higher

of these two values. This SFD measure was combined with other criteria to define three categories of population associated alleles (PAAs), each a sub-set of the previous. Initially zebu alleles were defined as those that are at higher frequency in the Indian population than in the European population, and taurine alleles are the converse. Category I PAAs include both zebu and taurine alleles that have a frequency of greater than 5% in the respective parental population. Alleles with a frequency lower than this were removed from the analysis. Category II PAAs are the subset of category I which are taurine alleles, i.e. those found at higher frequency within *B. taurus* populations. Category III PAAs are a more stringent selection of taurine associated alleles and are those found at a frequency of at least 20% in European populations and a SFD of greater than 40%. The average frequencies of categories II and III PAAs were calculated for each breed and these values were used as measures of *B. taurus* admixture.

2.2.3.c.i Support for selection of PAAs 1 -The Stepwise Mutation Model

The stepwise mutation model (SMM) of microsatellite evolution predicts that ancestry informative alleles should cluster in size in an allele length frequency histogram as they are assumed to have evolved from an ancestral allele in one or other species. It is considered, therefore, whether alleles designated as diagnostic of one or other ancestral group showed a tendency to cluster in size.

10,000 allele frequency distributions were simulated for each locus. Firstly the number of zebu and taurine assigned alleles were taken as in the observed data using the category I criteria. Secondly the order in which the alleles occurred on the length distribution was randomised. Every locus simulation was scored for allele intermingling as follows: An allele adjacency index (AAI) was calculated for each simulation by adding 1 when alleles designated to be from a single parental population were adjacent in size and subtracting 1

when alleles designated to be from different populations were adjacent. Here alleles are defined as adjacent if they are plus or minus one microsatellite repeat unit in size. This process was repeated for each of the loci and the total score for every 20 loci simulation was calculated and is plotted as a histogram in **Figure 2.3**. The AAI for the observed data at the 20 loci was calculated and compared to the distribution of simulations. This analysis was repeated excluding any allele that did not have a SFD of at least 40% (data not shown).

2.2.3.c.ii Support for selection of PAAs 2 – taurine alleles in the Near East

Allele frequencies for category II and category III taurine PAAs were calculated in seven Near Eastern cattle breeds using the MICROSATELLITE TOOLKIT (PARK 2001), among which a cline of ancestry has been described previously (LOFTUS *et al.* 1999). The frequencies of both of these categories of alleles were averaged to give an estimate of the fraction of taurine ancestry in these breeds. A Spearman rank correlation was carried out where actual rank is the average frequency of taurine alleles and expected rank is based on latitude.

2.2.3.d Genetic distance

Nei's standard genetic distance (D_S) (NEI 1972) was calculated using MICROSAT (MINCH *et al.* 1995) by the equation:

$$D_S = -\ln\left[\frac{J_{XY}}{\sqrt{J_X J_Y}}\right] \qquad \textbf{Equation 2.5}$$

where $J_X = \sum_{l=1}^{L}\sum_{i=1}^{n_l} p_{ix(l)}^2 \Big/ L$

and $$J_Y = \sum_{l=1}^{L} \sum_{i=1}^{n_l} p_{iy(l)}^2 \Big/ L$$

and $$J_{XY} = \sum_{l=1}^{L} \sum_{i=1}^{n_l} p_{ix(l)} p_{iy(l)} \Big/ L$$

in which $p_{ix(l)}$ and $p_{iy(l)}$ are the frequencies of the ith allele of the lth locus from populations X and Y, respectively. L is the number of loci, and n_l is the number of alleles at the lth locus (JIN and CHAKRABORTY 1994; TAKEZAKI and NEI 1996). The distance between each European breed and each Near Eastern and Indian breed was calculated separately and subsequently averaged.

D_{SW}, a distance measure accounting for the SMM (GOLDSTEIN *et al.* 1995a; SHRIVER *et al.* 1995), was not used, as it does may not perform significantly better than other distance measures for estimating phylogenies, even for microsatellite data (GOLDSTEIN *et al.* 1995b). D_S is considered to be one of the optimum measures when comparing divergence of populations (PAETKAU *et al.* 1997). Under both the infinite allele model and the stepwise mutation models of evolution, *Ds* is expected to increase linearly with time, if the mutation-drift balance is maintained throughout the evolutionary process. That is, the expected value *of Ds* = $\boldsymbol{2\mu t}$ where $\boldsymbol{\mu}$ is a mutation rate per locus per generation and $\boldsymbol{t}$ is the time measured in generations after divergence of the two populations (TAKEZAKI and NEI 1996). While most microsatellites are likely to mutate according the SMM (GOLDSTEIN *et al.* 1995a; JORDE *et al.* 1997), this pattern may be irregular for some loci and if loci are very polymorphic, the SMM may introduce excessive noise into data (BALLOUX and LUGON-MOULIN 2002; PAETKAU *et al.* 1997). The complex underlying mutation mode of

data analysed with the SMM can be difficult to interpret (DI RIENZO *et al.* 1994; ESTOUP *et al.* 1995b; WILSON and BALDING 1998). Some evidence indicates that compound or imperfect microsatellites may conform more closely to the infinite allele model (IAM) of mutation (GOLDSTEIN *et al.* 1995b), while further issues for analysis of both compound and perfect microsatellites include mutational biases or selective constraints on allelic size (GARZA *et al.* 1995; ORTI *et al.* 1997).

2.2.3.e *STRUCTURE* analysis

The STRUCTURE analysis program (PRITCHARD *et al.* 2000a) uses allele frequency data to characterise parental populations, assign individuals to these populations, detect admixed individuals and estimate individual admixture. The program uses a model-based clustering method that employs a Markov Chain to estimate the posterior distribution (q) of each individual's admixture co-efficient. The mean of this distribution has been denoted $\hat{q}$ and represents an estimate of the amount of an individual's genome that is derived from one of the inferred parental populations. No prior information was used to characterise the parental populations or to assign individuals to those populations. Two parental populations were inferred, following the guidelines recommended by Pritchard and colleagues. A burn-in of 50,000 repetitions and a run length of 1,000,000 were used.

2.3 RESULTS

2.3.1 DIVERSITY MEASURES

Diversity indices for the samples examined in this study are given in **Table 2.1**. Within the South Asian breeds, neither gene diversity nor MNA (mean number of alleles) showed a discernible geographical pattern. Gene diversity varied from 0.61 (Sahiwal) to 0.72 (Desi). In the five samples of comparable size, MNA values are lower in the two breeds of southern provenance, Ongole and Nellore than in Desi, Tharparkar and Red Sindhi. These measures may reflect differences in historical population size but may also be a result of higher levels of taurine admixture in the more northern samples (see below).

Table 2.1 *Table of breed names, codes and diversity values.*

Breed Name	*Code*	*N*	*Expected Heterozygosity*	*Exp Hz SD*	$\overline{A}$	*Sample size-adjusted* $\overline{A}$
Red Sindhi	RS	39	0.68	0.03	6.35	4.72
Tharparker	TH	35	0.61	0.03	5.20	3.95
Hariana	HA	10	0.71	0.03	5.15	5.01
Sahiwal	SA	10	0.60	0.05	4.40	4.27
Desi	DE	26	0.72	0.02	6.55	5.11
Nellore	NE	27	0.65	0.03	5.45	4.13
Ongole	ON	32	0.65	0.03	5.05	4.12

2.3.2 SELECTION OF POPULATION ASSOCIATED ALLELES (PAAs)

Population associated alleles have been used in many studies to infer admixture proportion. It is a method that is especially useful when there are large allele frequency differences between the two parental populations. *Bos indicus* and *Bos taurus* cattle are sufficiently divergent that there are many loci which show substantial allele frequency differences between the two species. These alleles have been used in previous admixture studies and produced results that have been supported by mitochondrial DNA and Y-chromosome

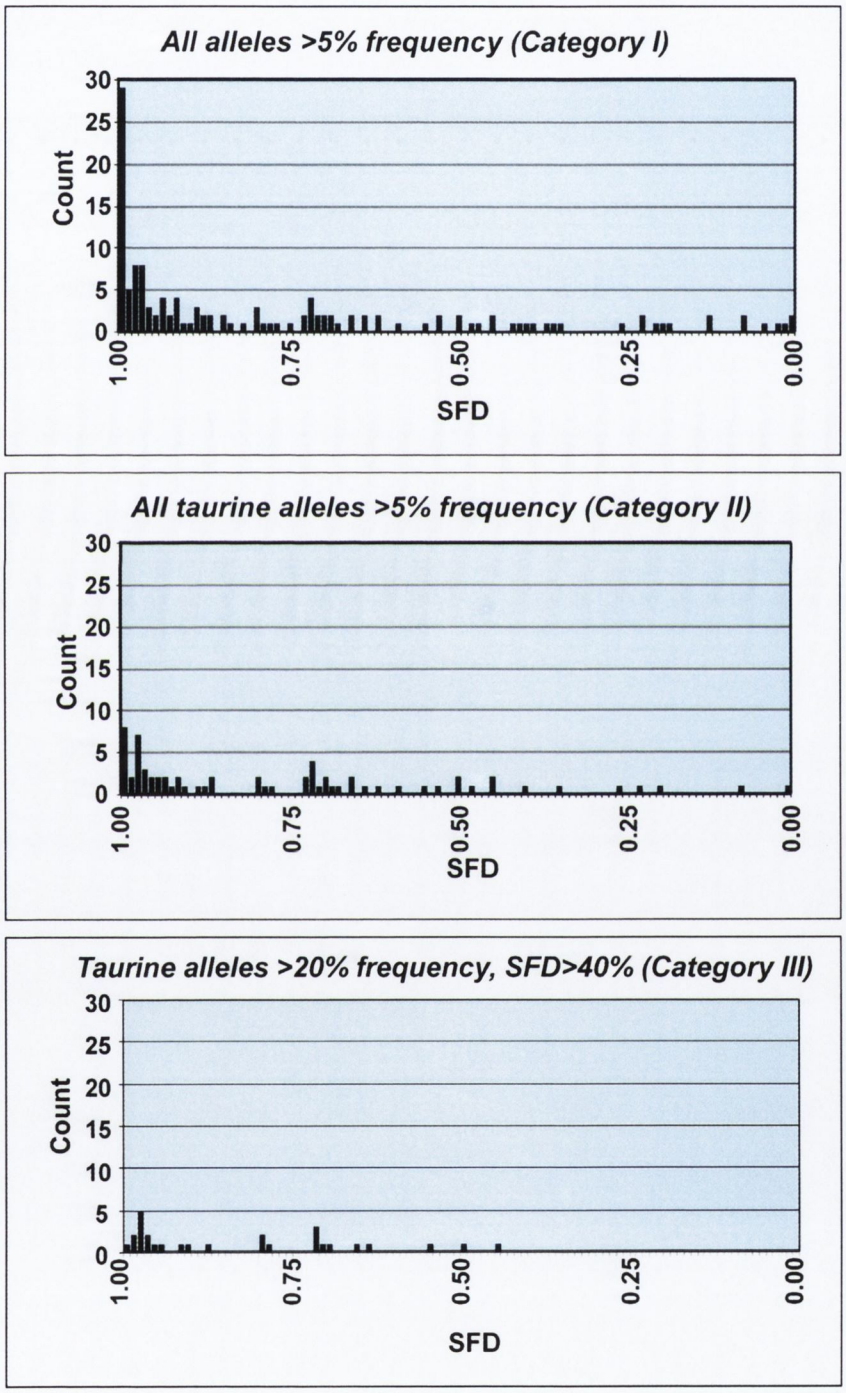
studies. However, the accuracy of such analysis is dependent on correct assignment of the parental populations.

Here alleles that are likely to be ancestry informative were selected by picking alleles that show a high frequency differential between a pooled population of European taurine cattle (n = 383) and the pooled data from each Indian sample (n =178). A pooled Indian sample was used to avoid designating one of the Indian breeds as a putative parental population. **Figure 2.2** shows histograms of scaled frequency differentials (SFD). These have been calculated for three allele categories.

- All alleles that occur at a frequency of 5% or greater in either the Indian or European pool.
- Taurine alleles that have a frequency of greater that 5%.
- Taurine alleles that have a frequency of greater than 20% and an SFD of greater than 40%.

Taurine alleles are those that have a greater frequency in European breeds than Indian breeds. The category I histogram demonstrates a predominance of alleles that have a high frequency differential between the two species. The alleles in the category III histogram that occur at 20% or more in the European population and have a SFD value greater than 40% are likely to be ancestry informative and have been utilised in the following analyses (**Figure 2.2**). A SFD value of 40% excludes few taurine alleles and similar levels of frequency differential have been found to be useful in other studies (McKeigue 1998; Parra *et al.* 1998).

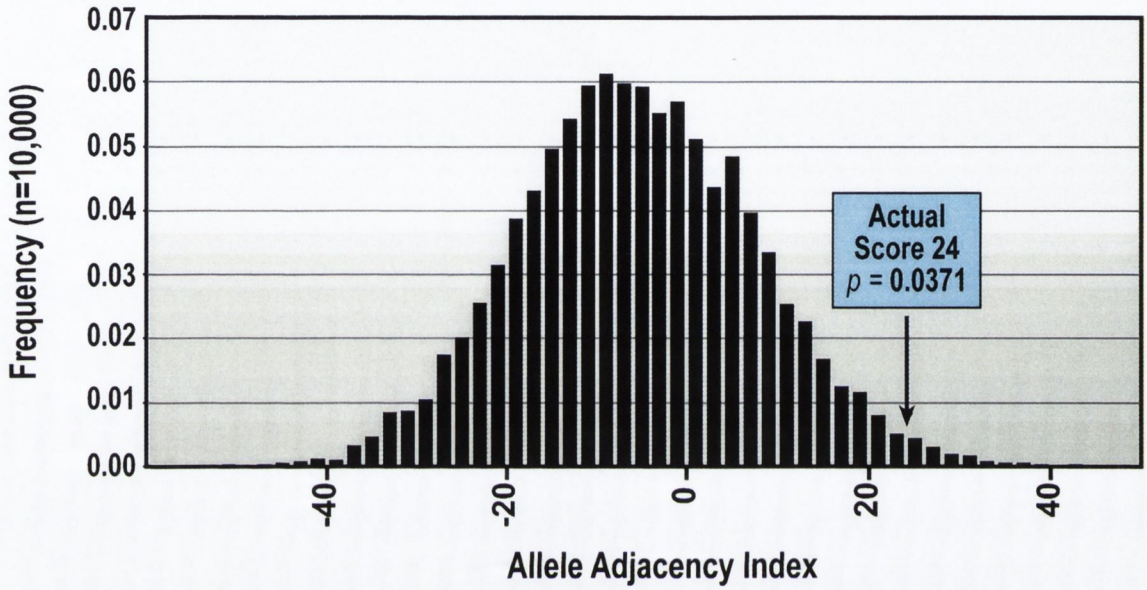
Figure 2.2 *Histogram of Scaled Frequency Differential (SFD) values for three groups of alleles typed in the European and Indian populations at 20 microsatellite loci. The SFD is a function of the allele frequency differential between the European sample and the entire Indian sample (see Material and Methods section). The first histogram (I) shows all alleles that occur at a frequency of at least 5% in either population. Many alleles in this group have a SFD of 1, as they are unique to one population. Taurine alleles are shown in the second and third histograms. These occur at a higher frequency in Europe than India and can be used to detect taurine admixture into Indian zebu cattle breeds. Each category is a subset of the previous one.*



2.3.2.a Support for PAA selection (1)

The stepwise model predicts that microsatellite alleles will mutate in discrete increments of repeat size (SHRIVER *et al.* 1993). If biologically meaningful PAAs have been selected alleles that are diagnostic of one lineage would be expected to cluster in size. It was necessary to ascertain whether this was the case for the loci in the present study as it acts as an independent control for our allele selection method. Therefore 10,000 sets of 20 loci were simulated with allele order on the allele length axis randomised and given an adjacency score. This allele adjacency index (AAI) is a measure of how often alleles from the same parental population were found to be adjacent in repeat unit count. The AAI null distribution and the score for the actual data are shown in **Figure 2.3**. This score is higher than 95% of the simulated scores indicating that the alleles from one lineage are significantly less mixed than would be expected by chance (p=0.0371). A more stringent allele designation method where alleles are required to have a SFD of at least 40% produces a highly significant result (p=0.0045). This indicates that ancestry informative alleles from one species have a significant non-random tendency towards similarity in size and thus supports common ancestral origins for selected taurine markers.

Figure 2.3 *Histogram of Allele Adjacency Index (AAI) for 20 simulated microsatellite loci. 10,000 sets of 20 loci were simulated in which the number of taurine and zebu alleles in each locus was held equal to the observed data but where the order in which they occur in the allele length distribution was randomised. The number of zebu and taurine alleles at the 20 simulated loci was determined by designating the alleles at the 20 study loci as either zebu or taurine using the category I allele criteria. An AAI was calculated for each simulation and these scores were plotted as a histogram, which is unimodal. An arrow shows the score for the real data.*



2.3.2.b Support for PAA selection (2)

To further test our selection method, the average frequencies of category II and III taurine PAAs in Near Eastern cattle populations were calculated. These populations were not used in the allele selection process and have been previously analysed by Loftus and colleagues (1999). The average taurine allele frequency for each breed was compared to breed longitude using a Spearman rank correlation. The *p*-value for this test is less than 0.001, supporting a decline in taurine influence (or increase in zebu influence) from west to east through the Near East region, as would be predicted from geography and prior analysis (LOFTUS *et al.* 1999).

2.3.3 GENETIC DISTANCE

2.3.3.a Application of PAAs to the Near East

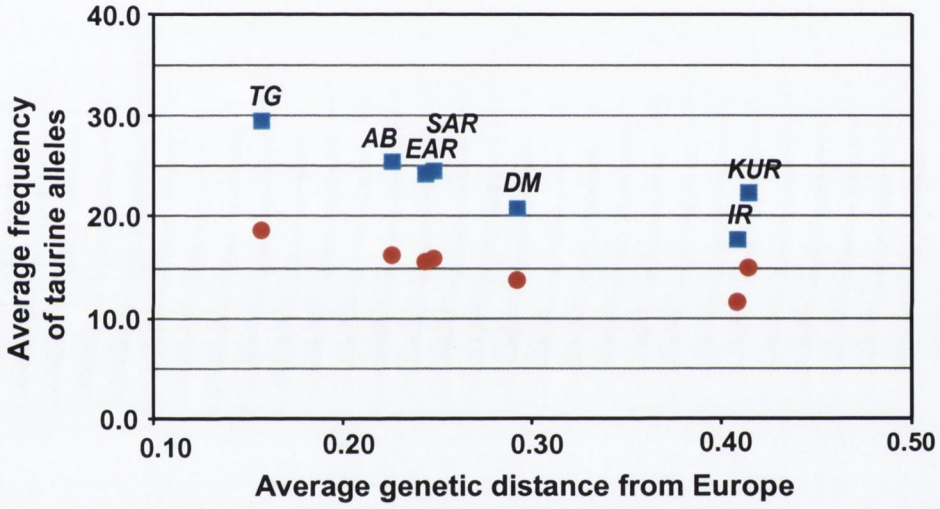
Admixture is known to truncate genetic distance. Hence, the average of Nei's standard genetic distance from the seven European breeds was taken for each Near Eastern population and plotted against PAAs frequency. The results for the two taurine PAAs categories are shown in **Figure 2.4a**. There is a negative correlation between the amount of taurine influence and the genetic distance from Europe (r^2 *(adj)* = 0.69, $p< 0.01$ for category III PAAs).

2.3.3.b Application of PAAs to India

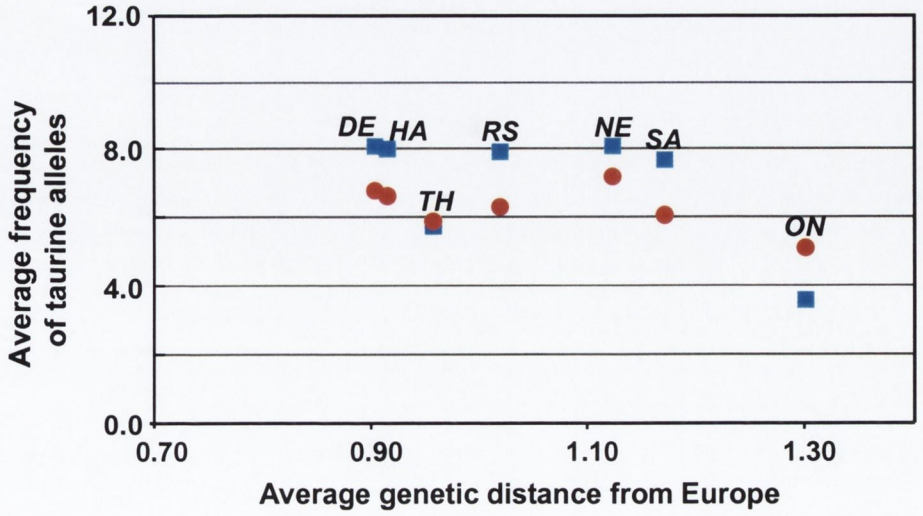
The average frequencies of our two categories of taurine PAAs in seven Indian cattle breeds were calculated. The southern-most Ongole breed has the lowest average frequency of both classes of taurine alleles. The average taurine allele frequency for each breed was used in a Spearman rank correlation using geographical latitude as ranking parameter. This test does not support any simple cline throughout the data. The average of Nei's standard genetic distance from the seven European breeds was plotted against the average taurine allele frequency of category II and category III alleles (**Figure 2.4b**). From both these data treatments the Ongole breed appears to be the least introgressed of the Indian breeds.

Figure 2.4 *Scatter plots of average of Nei's standard Genetic Distance from Europe vs. the average frequency of category II and III taurine alleles. Nei's genetic distance was calculated between each of the (a) Near Eastern and (b) Indian breeds and each European breed. The distance to European Bos taurus was averaged for each of the Near Eastern and Indian breeds. This value was plotted against the average frequency of category II and III alleles. The circles represent the category II alleles and the squares the category III alleles. The category II allele criteria requires alleles to occur at a minimum frequency of 5% in Europe while the category III alleles require a European frequency of 20%. The first plot shows a cline of taurine influence from west to east in the Near Eastern cattle breeds. The regression for this plot is significant ($p < 0.01$). The second plot shows the Indian breeds. The regression for this line is not significant but it is clear that Ongole has the lowest frequency of category II and III taurine alleles.*

a. Scatter plot of average Nei's Distance from Europe vs. average frequency of taurine alleles in 7 Middle Eastern breeds



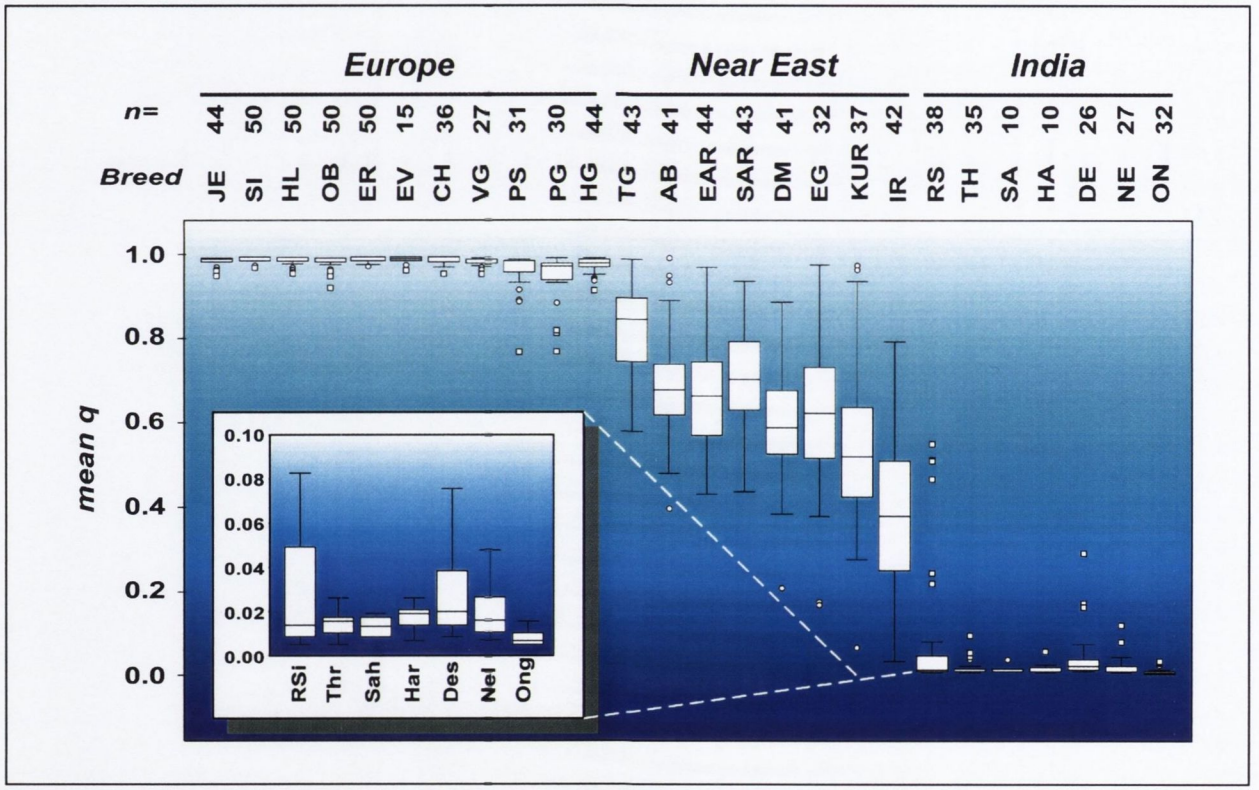
b. Scatter plot of average Nei's Distance from Europe vs. average frequency of taurine alleles in 7 Indian breeds



2.3.4 STRUCTURE ANALYSIS

The STRUCTURE integrated Bayesian approach both infers parental population allele frequencies and estimates the mean of the posterior distribution of each individual's admixture co-efficient ($\hat{q}$) (PRITCHARD *et al.* 2000a). Two parental populations were assumed for the admixture analysis. This supposition is supported by the inferred allele frequencies for the two parental populations, between which the average SFD is 0.7. The posterior distributions of admixture proportions for all individuals are not uniform and some have very narrow distributions that do not span either 0 or 1. Clearly one of the two inferred populations is characterised by European *B. taurus* individuals and the other by Indian *B. indicus* cattle. These have $\hat{q}$ values between 0.9-1.0 (taurine) and 0.0-0.1 (zebu). Beaumont et al. (2001) used these values of $\hat{q}$ to differentiate between parental ($\hat{q}$ = 0-0.1 and 0.9-1.0) and hybrid populations ($\hat{q}$ = 0.1-0.9). It can be argued that this output reflects the ancestral divergence between *B. taurus* and *B. indicus*. The values of $\hat{q}$ for each individual have been plotted according to breed in the box plot in **Figure 2.5**. The breeds with the highest and lowest median $\hat{q}$ values are the Desi (0.022) and the Ongole (0.009) breeds, respectively. Values of $\hat{q}$ range from only 0.006 to 0.03 in Ongole samples, while $\hat{q}$ values from 0.006 to 0.552 are observed in the Red Sindhi breed of Northern India.

Figure 2.5 *Box plot of mean individual admixture coefficient ($\hat{q}$) for each individual grouped by breed. Two parental populations were inferred by the STRUCTURE program without using prior sample information. One of these populations is characterised by European taurine and the mean of the posterior distribution of likely ancestry from this population was plotted for each individual. The hybrid status of the Near Eastern breeds is demonstrated by their position between Europe and India and the wide data range for each breed. The scale of the box plot makes it difficult to ascertain any difference between the seven Indian breeds but if this section of the box plot is magnified (see inset) it is clear that the Ongole samples have the lowest mean values of* $\boldsymbol{q}$. *The box represents the interquartile range, which contains 50% of the values and line across the box indicates the median. The whiskers are the lines that extend from the box to the highest and lowest values, excluding outliers and extreme values. Outliers, which are represented by circles, are values that are 1.5 to 3 box lengths from the upper of lower edge of the box. Extreme values, which are greater then 3 box lengths from the upper or lower edge, are represented by squares.*



2.4 DISCUSSION

2.4.1 PATTERNS OF GENETIC VARIABILITY IN THE BREEDS SURVEYED

An older source population will typically display greater genetic diversity than a population derived more recently from it (JORDE *et al.* 1998). However, within the South Asian breeds, neither gene diversity (heterozygosity) nor MNA showed a discernible geographical pattern which could point to a likely region of origin (**Table 2.1**). In the five samples of comparable size, MNA values are lower in the two breeds of southern provenance, Ongole and Nellore than in Desi, Tharparkar and Red Sindhi.

2.4.2 INTROGRESSION OF *B. taurus*

Microsatellite allele size, as well as frequency, is a source of information about ancestral relationship. One feature of this property in cattle is that with a proportion of markers it seems that *B. indicus* and *B. taurus* populations display separate allele length distributions. It has been argued previously that through identification of allele classes which were close to fixation in pure zebu, and absent or near-absent in breeds without their ancestral influence, the existence of geographical clines stretching from Europe to India can be shown (LOFTUS *et al.* 1999; MACHUGH *et al.* 1997). Here, a symmetrical approach is taken and the sets of alleles in loci which comprise the ancestral *B. taurus* distribution are estimated.

A potential obstacle to this approach is selection of the parental population. The selection of private alleles could be biased by the incorrect assignment of parental or hybrid status. The difficulties in estimating population and especially individual admixture when gene frequencies in the parental population before the admixture event are unknown have been discussed previously (BERTORELLE and EXCOFFIER 1998; ESTOUP *et al.* 1999). Incorrect

assignment of parental populations may lead to an over-estimation of levels of admixture. In many cases either temporal or spatial phenomena will mean that it is impossible to sample from the original parental populations that contributed to the admixture event. The issue of selecting ancestry informative alleles is further confounded by the fact that some parental populations may have overlapping allele frequency distributions and there may be very limited numbers of private alleles that are found in only one or the other contributing population. In cases such as this researchers are forced to select a reasonable frequency differential to characterise population associated alleles (PARRA *et al.* 1998; PFAFF *et al.* 2001).

In the current study three alternative methods were used to overcome these difficulties. A scaled frequency differential between large pooled breed groups of alternately *B. taurus* and *B. indicus* cattle was used to conservatively assign alleles to parental populations. Examination of these illustrates relative influences of parental populations in hybrid populations. The genetic distance of each of the seven Indian breeds from all European breeds using Nei's standard method was also calculated. Finally, the STRUCTURE program, which does not require parental populations to be pre-specified and which uses a Bayesian approach to define parental populations and to assess individuals, was employed.

The discriminating power of population-associated alleles is implied when all alleles with a frequency greater than 5% are examined and show a high proportion of either taurine or zebu associated alleles (**Figure. 2.2**, Category I). Also, alleles at any locus which are associated with one subspecies are more likely to be similar in repeat number; implying a genealogical relationship which supports their utility as indicators of ancestry (**Figure 2.3**).

The average frequencies of categories II and III alleles were calculated for seven Near Eastern breeds and plotted against Nei's standard genetic distance from Europe. There is a significant correlation between the two measures and a Spearman rank correlation using longitude as a ranking variable supports a gradation of average taurine allele frequency from west to east with a *p*-value of less than 0.001. This supports previous findings about the relationship between these breeds and is a reflection of the former herding of zebu cattle alongside taurine cattle in places such as Mesopotamia and Iran (LOFTUS *et al.* 1999). As these populations were not used in the initial allele designation this pattern provides an independent test for whether the alleles are ancestry informative.

When the frequency of these alleles in Indian breeds was examined, *B. taurus* alleles were found in each population surveyed. While it is difficult to differentiate between the amounts of taurine influence in the Indian breeds, the PAA analysis suggests that the southern Ongole breed may have experienced the least taurine influence (**Figure. 2.4b**).

The STRUCTURE program confirms the presence of *B. taurus* influence in the Indian sub-continent. The algorithm presumes that the original ancestral parental populations were in Hardy-Weinberg equilibrium, with the markers used in complete linkage equilibrium, and that the two parental populations contributed alleles to the hybrid population as independent draws. This model fairly accurately represents the original domesticated cattle populations of the India and the Near East, as it seems likely that they were separated for millennia prior to any contact. STRUCTURE has the important advantage of modelling the ancestral parental population allele frequency distributions without *a priori* selection of these populations. It also allows admixed individuals to be included in the process of characterising these ancestral distributions. This approach also selects the Ongole as

having the lowest likelihood of taurine influence (**Figure. 2.5**). There is also a highly significant regression between the mean value of $\hat{q}$ for Indian breeds and their average genetic distance from Europe ($p<0.0001$, $r^2(adj) = 0.898$). Therefore the data indicate some ancestral influence from taurine cattle, probably of Near Eastern origin. This is also suggested by the fact that highest values of individual admixture are in the Red Sindhi breed, which is geographically closest to the Near East.

The majority of Near Eastern individuals have posterior distributions with means between 0.1 and 0.9; a result of their documented hybrid status (LOFTUS *et al.* 1999). Both Beaumont (2001) and Hansen (2001) found that most hybrids with intermediate values of $\hat{q}$ had 90% posterior intervals (PI) that ranged from 0.0 to 1 making it difficult to reliably identify admixed individuals. By contrast, $\hat{q}$ here is between 0.1 and 0.9 for 33.6% of the sample in the present study and of these hybrid individuals 67% have 90% PI that are between 0.1 and 0.9. As the two previously mentioned studies used only nine microsatellite loci and 20 are included in the current study, the suggestion that an increased number of markers will enhance resolution (HANSEN *et al.* 2001) is supported by our results.

The extent of taurine influence in the some of the India samples is somewhat surprising given the distinctly zebu morphology of all of the breeds surveyed. However, this is part of a broad geographic pattern of introgression with a cline in admixture stretching through the fringes of Europe, Anatolia and the Middle East towards India. The presence of taurine alleles in India may be explained by either an ancient or more recent introduction of *B. taurus* into the Indian sub-continent. A significant primeval introduction associated with

the inception of pastoralism in the region would be expected to leave a detectable genetic signature in modern populations. The lack of taurine influence in the southern Ongole breed suggests that the observed taurine introgression is more likely to be the result of post-domestication introgression, or that any original *B. taurus* input was minor, and has been subsequently saturated by indigenous sources. It is worth noting that there may have been attempts to improve some modern Indian cattle by very recent importation of taurine individuals (FELIUS 1995). However, in the current study, the unimproved, nondescript Desi cattle act as a control for this possible modern influence.

Locally domesticated zebu cattle were a prominent component of the Harappan civilization of the Indus valley in the third millennium BC, as indicated by faunal and artistic evidence (ALLCHIN 1969; MEADOW 1996; ZEUNER 1963). By this period, however, some humpless taurine cattle were also present, as they are depicted on some Harappan stamp seals. In addition evidence, both from ancient art as well as some bone assemblages, indicates that zebu were present in Sistan, Iran and further west in Mesopotamia and even Kamid el-Loz in Lebanon during the Bronze Age, i.e. 3000 BC-1000 BC (BÖKÖYNI 1990; BÖKÖYNI 1997; CLASON 1978; MEADOW 1987). This evidence for the spread of zebu westwards to the Near East and the spread of taurine cattle eastwards to the Indus valley supports that the introgression between taurine and zebu cattle indicated by the genetic data could have began by the third millennium BC and has continued in varying degrees ever since. Archaeological evidence for long-distance trade indicates there has been continuous contact between the Indus region and Mesopotamia since the fourth millennium BC (ALLCHIN and ALLCHIN 1982; CHAKRABARTI 1990; LAMBERG-KARLOVSKY 1989).

It was also through this same region (Baluchistan) that a series of folk movements of peoples from the north-western periphery of South Asia are postulated to have occurred during the course of and after the Harappan civilization (ALLCHIN and ALLCHIN 1982). The scale of these migrations, and whether or not they are to be connected with the diffusion of Indo-European languages as traditionally assumed (PARPOLA 1994), remains highly controversial (CHAKRABARTI 1997; RENFREW 1987; SHAFFER 1984). Nevertheless, such human movements, presumably of agropastoral groups, would have contributed to the further introgression of taurine genes into zebu populations.

At this period, or perhaps in preceding millennia, flax and a number of pulses, including peas, lentils, grasspea and chickpea, spread from west to east. During the third millennium some subsistence species spread westwards, including sesame (CHARLES 1993), the chicken (BÖKÖYNI 1990), and water buffalos (ZEUNER 1963). Therefore, the dispersal of *Bos taurus* cattle into South Asia, and the counterflow of some *Bos indicus* into Mesopotamia and the Levant can be seen as part of a larger pattern of the prehistoric transfer of subsistence species between these two primary centres of agricultural innovation (FULLER 2001).

CHAPTER 3

A population genetic survey of West African cattle populations

3.1 INTRODUCTION

As outlined in **Chapter 1**, the original cattle of Africa were *B. taurus*. They were the descendants of aurochs that were domesticated in either the Near East or on the African continent itself (BRADLEY *et al.* 1996; CLUTTON-BROCK 1989; EPSTEIN and MASON 1984; HANOTTE *et al.* 2002). Since that time waves of migrations of indicine cattle have changed the genetic landscape of African cattle completely (HANOTTE *et al.* 2002; HANOTTE *et al.* 2000; LOFTUS *et al.* 1994b; MACHUGH *et al.* 1997). Molecular genetic surveys of autosomal microsatellite markers reveal that most extant cattle breeds in the Sahel region have had substantial zebu genetic input, except for a small number of relic African taurine populations located further south in the humid forested areas (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). However, the finding that all African cattle surveyed to date have *B. taurus* type mitochondrial DNA reveals their original ancestry (TROY *et al.* 2001).

Payne and Hodges (1997) define three indigenous West African cattle types: longhorn N'Dama of the far west forest savannah, dwarf West African shorthorn such as Baoule, and Kuri from the wetland areas of Lake Chad. It is feasible that these African taurine populations are the descendants of the original *B. taurus* shorthorns that were domesticated in North Africa, possibly after initial introductions of taurine cattle from the Fertile Crescent. Available archaeological evidence suggests that pastoralists herding shorthorn cattle had migrated to and penetrated West African forests by about 4000 BP (MACDONALD and HUTTON MACDONALD 2000; MARSHALL and HILDEBRAND 2002).

N'Dama are a longhorned, humpless breed centred on the Fouta Djallon plateau of Guinea, which can inhabit areas where the tsetse fly is endemic, as a result of inherent trypanotolerance. Shorthorned West African *B. taurus* are also trypanotolerant, and it is

likely that all these cattle have been in West Africa for a considerable period, which would have enabled the acquisition of the enhanced tolerance of trypanosomiasis (PAYNE and HODGES 1997). They may have entered the region as early as 7000 BP, but were not well established in the river basins until around 3000 years later (MACDONALD and HUTTON MACDONALD 2000). Kuri cattle are large longhorned taurines, which inhabit the wetland environments around Lake Chad. This breed is not trypanotolerant (EPSTEIN 1971).

B. indicus cattle are newcomers to the region and do not exhibit any trypanotolerant traits. As described in **Chapter 1**, zebu cattle were first introduced to the east of the African continent c. 2000-3000 BP (CLUTTON-BROCK 1989) and major influxes of these animals followed local Arabian contacts or the establishment of Indian Ocean trade (HANOTTE *et al.* 2002). The earliest representations of humped cattle in West Africa date from around 1000 BP, but there is very little evidence for when *B. indicus* first entered this part of the continent in significant numbers (MACDONALD and HUTTON MACDONALD 2000).

In West Africa, the extent of the zebu component of each breed has been found to vary according to the climate and associated trypanosomiasis incidence (MACHUGH 1996; MEGHEN *et al.* 1994). While indicine cattle are better adapted to the increasingly dry areas of West Africa than the taurine breeds, the local trypanotolerant taurine N'Dama and shorthorn cattle do not thrive as successfully in arid areas. Therefore populations further north (e.g. in Mauritania) and east (e.g. in Nigeria) have a substantial indicine genetic component, whereas populations to the southwest (e.g. in Guinea) have maintained most of their taurine genetic integrity (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). There is a hybrid zone between the two cattle types, which is maintained along the margins of the humid areas where the tsetse fly is endemic (**Figure 3.1**).

This study represents the most extensive to date in this region. Previous analysis revealed the gradient of zebu admixture using populations associated alleles and principal components (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). A variety of methods, including multivariate analysis, traditional admixture analysis and newly developed Bayesian methods, are employed to get an accurate picture of the current and historical patterns of hybridisation and gene flow in this part of the African continent.

3.2 MATERIALS AND METHODS

3.2.1 SAMPLE COLLECTION AND GENOTYPING

Blood samples were collected on several sampling trips to West Africa ensuring, when possible, that the animals selected were unrelated and considered typical of the breed. In some cases (e.g. N'Dama), several different populations of what would traditionally be classed as one breed were sampled, as it is likely that the same breed sampled in different countries or regions may have different histories of introgression. The concept of breed is not one that is simple to define in an African context, as groups of cattle are often named after the people by whom they are kept. For the purposes of the following analyses, each sample population will be referred to as a breed. DNA from blood was extracted and samples were screened with 20 microsatellite loci by Ciaran Meghan, using the procedures outlined elsewhere (MACHUGH *et al.* 1997). The data were combined with previously published data from other cattle breeds (MACHUGH *et al.* 1997). Sampling locations for African breeds are shown in **Figure 3.1**. Other relevant sample information is given in **Table 3.1**. Populations giving data that are novel to the current study are highlighted with an asterisk.

Figure 3.1 *Map of sampling area (reproduced from MEGHAN et al., 1994) Sampling locations for taurine-type breeds are shown as circles and for indicine breeds are shown as squares. The area of West Africa where the tsetse fly is known to be endemic is highlighted*

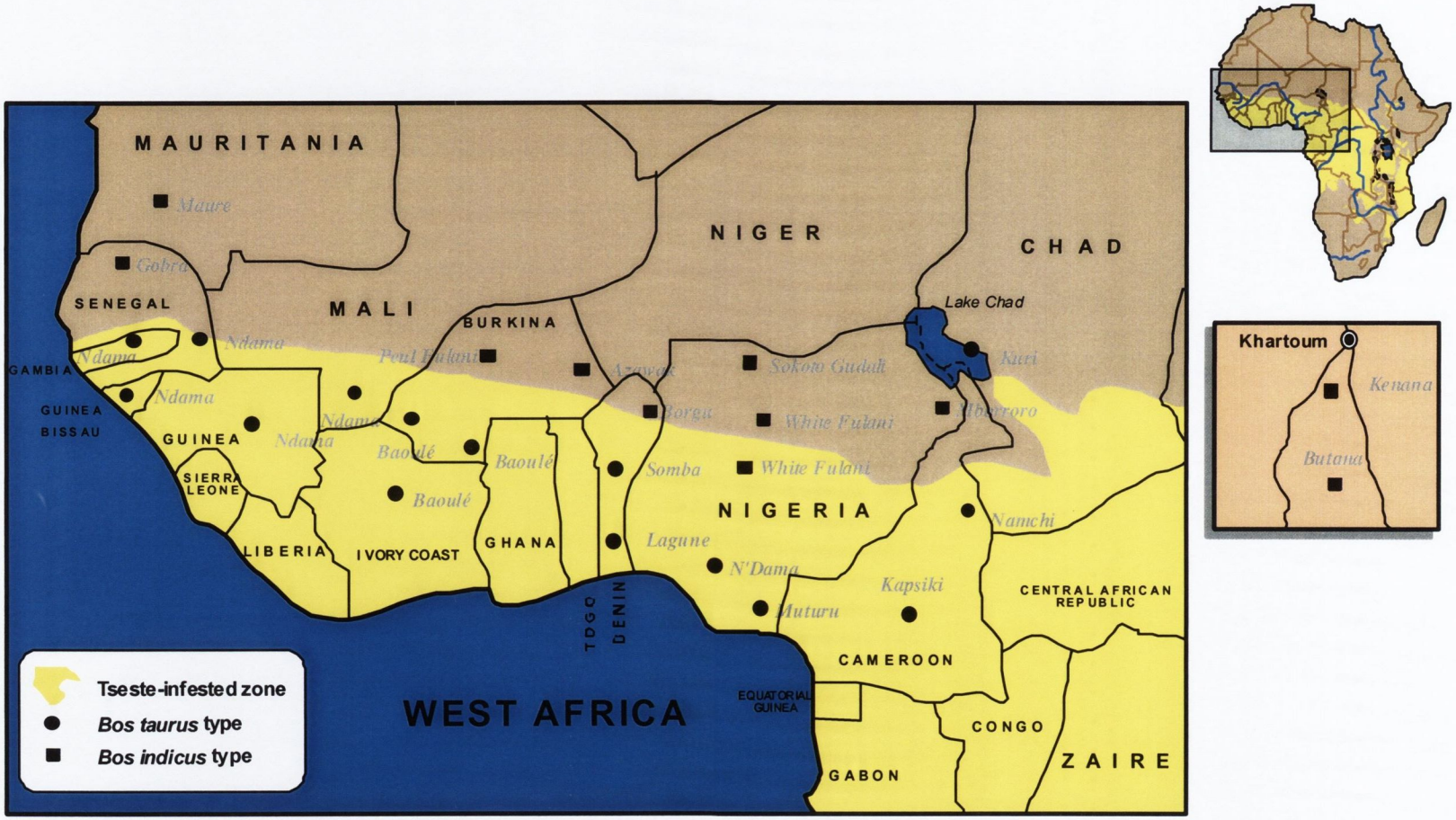


Table 3.1. *Breed information for the 36 breeds included in the study.*

Continent	Population	Breed origin	Code	Sample size	$\hat{H}$	Size-adjusted $\overline{A}$
Europe	*Aberdeen Angus*	UK	**AN**	33	0.4773	3.047
	Charolais	France	**CH**	36	0.525	3.582
	Friesian	Netherlands	**FR**	40	0.5513	3.712
	Hereford	UK	**HE**	34	0.4647	3.067
	Jersey	UK	**JE**	34	0.4397	2.856
	Kerry	Ireland	**KR**	40	0.4788	2.846
	Simmenthal	Switzerland	**SI**	36	0.4986	3.448
Africa	*Baoule**	Cote d'Ivoire	**BC**	46	0.5374	3.885
	*Baoule**	Burkina Faso 1	**BG**	56	0.4624	3.460
	*Baoule**	Burkina Faso 2	**BL**	50	0.5114	3.772
	Butana	Sudan	**BU**	24	0.6521	3.648
	*Fulani**	Nigeria	**FU**	24	0.6104	3.955
	Gobra	Senegal	**GO**	59	0.6331	4.421
	Kenana	Sudan	**KE**	38	0.6434	3.904
	Maure	Mauritania	**MA**	55	0.6309	4.589
	N'Dama	Guinea Bissau	**NB**	54	0.538	3.578
	N'Dama	Guinea	**NGu**	63	0.5127	3.308
	N'Dama	The Gambia	**NGa**	58	0.5836	4.102
	N'Dama	Nigeria	**NN**	19	0.5289	3.820
	N'Dama	Senegal	**NS**	48	0.5323	3.910
	*N'Dama**	Mali	**NM**	44	0.5331	3.878
	*Kapsiki**	Cameroon	**KA**	60	0.5912	4.192
	*Kuri**	Chad	**KU**	52	0.6154	4.374
	*Laguna**	Benin	**LA**	60	0.4517	3.327
	*Muturu**	Nigeria	**MU**	28	0.4075	3.275
	*Namchi**	Cameroon	**NA**	54	0.5196	3.818
	*Somba**	Toga	**SO**	60	0.5157	3.639
	*Azawak**	Burkina Faso	**AZ**	29	0.6023	3.823
	*Borgou**	Benin	**BO**	60	0.6434	4.352
	*Mborroro**	Nigeria	**MB**	47	0.6016	4.177
	*Peul Fulani**	Burkina Faso	**PF**	79	0.6061	4.352
	*Sokoto Gudali**	Nigeria	**SG**	47	0.6048	4.180
	White Fulani	Nigeria	**WF**	30	0.595	3.980
India	*Hariana*	India	**HA**	10	0.55	3.595
	Sahiwal	India	**SA**	13	0.5423	3.584
	Tharparker	India	**TH**	10	0.635	3.294

3.2.2 DIVERSITY MEASURES

Allele frequencies for all loci were determined directly by standard gene counting and are tabulated in **Appendices A3 and A4**. Average unbiased estimates of gene diversity, also denominated expected heterozygosity ($\hat{H}$), with associated standard errors were computed for all locus/population combinations according to Nei (1987) as described in Section 2.2.3.b. Sample size-corrected mean number of alleles per locus ($\overline{A}$) was also calculated as an index of sample diversity using the MICROSATELLITE TOOLKIT as outlined in Section 2.2.3.a (PARK 2001). The values were averaged over all loci and sampling with replacement was used.

3.2.3 ADMIXTURE ANALYSIS

3.2.3.a Least-squares regression

mR (ROBERTS and HIORNS 1965) was calculated using the ADMIX *v.* 1.0 program (www2.unife.it/genetica/Giorgio/Giorgio.html). The data from all of the European breeds (n=253) and the pooled data from each Indian sample (n=33) were used as parental populations for this analysis. These populations were also used by Hanotte and colleagues as putative ancestral populations (2002). It may be argued that African *B. taurus* breeds would represent a more appropriate taurine parental population, and an advantage of the Bayesian analysis presented below is that it does not use *a priori* identified parental populations. *mR* (range 0.0 to 1.0) was also used to assign each breed as either taurine or indicine. Breeds with a *mR* value greater than 0.5 were denominated as African taurine and those with values less than 0.5 were defined as African indicine. Taurine breeds are shown as circles on the sample map (**Figure 3.1**) while indicine breeds are shown as squares.

3.2.3.b *Population-associated alleles*

B. indicus associated alleles were selected by comparing allele frequencies from the pooled European taurine breed data with pooled data from all Indian zebu samples. Using a symmetrical approach to that used to define taurine PAAs in **Chapter 2** (Section 2.2.3.c), scaled frequency differential (SFD) was calculated for each allele by the formula

$$\frac{\left|P^E - P^I\right|}{\max(P^E, P^I)} \qquad \textbf{Equation 3.1}$$

where $\boldsymbol{P^E}$ is the average frequency of the allele in the European taurine breeds and $\boldsymbol{P^I}$ is the average frequency in the Indian indicine breeds assayed. The higher of these two values is denoted by $\boldsymbol{max\ (P^E,P^I)}$. This SFD measure was used to define zebu-associated alleles (PAAs). For these analyses alleles which occur at a frequency of at least 20% in the Indian *B. indicus* sample and which have a SFD value of at least 40% were used. These are equivalent to category III alleles which are defined in Section 2.2.3.c (KUMAR *et al.* 2003). The average frequency of these category III zebu PAAs was calculated for each breed and these values were used as measures of admixture.

3.2.3.c *Principal Components and Multi-Dimensional Scaling*

Principal components analysis was carried out with the GENSTAT package and multidimensional scaling (MDS) was carried out using F_{ST} values calculated by the ARLEQUIN software (SCHNEIDER *et al.* 2000). These procedures simplify multivariate data with minimum loss of information (CAVALLI-SFORZA *et al.* 1993). Admixed populations may contradict principles of phylogeny reconstruction, so multivariate analyses are especially useful for such groups (MOAZAMI-GOUDARZI and LALOE 2002). F_{ST} is a widely used and convenient measure of genetic differentiation among subpopulations (SLATKIN

1995; WEIR and COCKERHAM 1984; WEIR and HILL 2002; WRIGHT 1969). It is defined by the equation:

$$F_{ST} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}} \qquad \textbf{Equation 3.2}$$

where F_{IT} and F_{IS} are measures of deviation from Hardy-Weinberg proportions in the total population and within subpopulations respectively (HEDRICK 2000; WRIGHT 1951; WRIGHT 1965). Positive values for these parameters indicate a deficiency of heterozygotes, while an excess of heterozygotes is represented by negative values. F_{ST} may also be calculated for multiple loci by

$$F_{ST} = \frac{\overline{H}_T - \overline{H}_S}{\overline{H}_T} \qquad \textbf{Equation 3.3}$$

where $\overline{H_S}$ is the average expected heterozygosity within subpopulations over loci and $\overline{H_T}$ is the average of the expected heterozygosity in the total population over loci (HARTL and CLARK 1997; NEI 1987). For the reasons described in Section 2.2.3.d, R_{ST}, (SLATKIN 1995), an analogous measure to F_{ST} which has been optimised for loci evolving according to the assumptions of the SMM, was not used. Moreover, high variance is often associated with R_{ST}, and this can affect population differentiation estimates, even for markers that correspond perfectly to the stepwise mutation model (SMM) (BALLOUX *et al.* 2000). F_{ST} values calculated from microsatellite data have been shown to correspond to spatial population distributions (KANTANEN *et al.* 2000; MALASPINA *et al.* 2000).

Multidimensional scaling (MDS) is a statistical technique that can be used to illustrate population divergences based on a matrix of inter-population genetic distances (MANLY 1986). This is accomplished by assigning observations to specific locations in a two-dimensional conceptual space, such that the distances between points in the space match the given dissimilarities as closely as possible. In many cases, the dimensions of this conceptual space can be interpreted and used to further understand your data. Unlike phylogenetic trees, MDS does not impose ancestral status upon the populations being compared, but instead demonstrates the relationships among populations, relative to each other, based on the genetic distances calculated between them. MDS attempts to discover any groupings in a set of distance measures between populations. In addition to this, each dimension may illuminate a discrete genetic trend. MDS was implemented using SPSS (*v.* 11.0, SPSS Inc. Chicago, IL)

3.2.3.d Genetic Distance

Pairwise Nei's standard genetic distances (D_S) were calculated between breeds using **Equation 2.5** as implemented by the MICROSAT program (MINCH *et al.* 1995). Distances between the African breeds were used to construct a neighbor-joining tree using the PHYLIP (*v.* 3.572c) package (FELSENSTEIN 2001). The method is based on finding pairs of operational taxonomic units that minimise the total branch length at every stage of clustering (SAITOU and NEI 1987). The average genetic distance between each of the African breeds and the pooled Indian samples was calculated and used as a measure of *B. indicus* admixture.

3.2.3.e STRUCTURE analysis

The STRUCTURE analysis program (PRITCHARD *et al.* 2000a) uses allele frequency data to characterise parental populations, assign individuals to these populations, detect admixed

individuals and estimate individual admixture. The program uses a model-based clustering method that employs a Monte Carlo Markov Chain (MCMC) to estimate the posterior distribution (q) of each individual's admixture co-efficient. The mean of this distribution has been denoted $\hat{q}$ and represents an estimate of the amount of an individual's genome that is derived from one of the inferred parental populations. No prior information was used to characterise the parental populations or to assign individuals to those populations. After running multiple simulations to optimise model parameters three parental populations were inferred. Log likelihood values (**Figure 3.6**) and previous evidence (HANOTTE *et al.* 2002; HANOTTE *et al.* 2000; MACHUGH *et al.* 1997) suggested that a three-way admixture model was appropriate. A burn-in of 100,000 repetitions and a run length of 1,000,000 were used. Contour maps of admixture coefficients and $\overline{A}$ values were constructed using ARCVIEW GIS *v.* 3.2 (Environmental Systems Research Institute, ESRI, Redlands, CA). A projection of 1:55,000,000 and equal interval contours were used to visualise the data. Maps were produced using the Adobe ILLUSTRATOR package.

3.3 RESULTS

3.3.1 DIVERSITY MEASURES

Diversity indices for the samples examined in this study are given in **Table 3.1**. Within the West African breeds, there is a discernible pattern where breeds such as Gobra, Maure, Kuri and others located around the perimeter of the tsetse zone have considerably higher values of $\overline{A}$ than most other breeds within the zone or in East Africa. Within Africa, Guinean N'Dama has the lowest levels of diversity, by this measure. It is noteworthy that the diversity levels in Europe are lower than most of the African breeds. $\overline{A}$ is shown in a contour map in **Figure 3.2**. This measure will be increased by admixture and seems to be is highest in those breeds with a history of *B. indicus* introgression. Diversity levels are relatively low in the East African zebu breeds as the taurine component of their genome is relatively small.

3.3.2 MULTIDIMENSIONAL SCALING ANALYSIS AND GENETIC DISTANCE.

The multidimensional scaling (MDS) of pairwise F_{ST} genetic distances clearly separates the *B. taurus* and *B indicus* lineages in its first dimension (**Figure 3.3**). This axis accounts for 94.3% of the total variation in the data. Groups of breeds that can be distinguished are bounded by ovals. The positioning of the groups along axis one corresponds to the extent of their *B. indicus* component. They are ordered as follows: the East African zebu group has experienced the greatest zebu influence, followed by West African zebu, than Kuri, Kapsiki and Borgou, and finally the West African taurine groups, Namchi and Muturu. The second axis of variation (3%) seems to group breeds from similar geographic locations. In fact, there is separation between the East and West along this second axis of variance. Blue shading around the breed code designates breeds that are found in the Western part of West Africa. Those from the more eastern part of the region are shown by

grey shading behind the breed code. N'Dama and Baoule, from the humid forested western-most part of the region, form a recognisable cluster. There is also grouping of the shorthorn breeds from Benin, and of other taurine breeds from Cameroon and Nigeria. Blue and grey shading also indicate some east/west geographical separation of the zebu breeds along this axis.

3.3.3 GENETIC DISTANCE

A neighbor-joining tree constructed from Nei's standard genetic distance, (**Figure 3.4**) including European, African and Indian cattle breeds, displays truncated branch lengths for the hybrid breeds. Association of breeds in groups is evident from this analysis. Indian *B. indicus* breeds form a cluster, which is located closest to East African zebu. West African zebu breeds and hybrids are intermediate to East African *B. indicus* and West African taurine breeds. European *B. taurus* are found beside African taurines, at the opposite extreme of the tree. Short branch lengths could be the results of different rates of evolution along different lineages, but in this case are more likely to be indicative of admixture (MOUNTAIN *et al.* 1992). Short or absent branches off the main trunk of the tree are especially evident for most of the zebu and hybrid breeds of West Africa, but are also a feature of some West African taurine breeds.

Figure 3.2 *Contour map of size adjusted mean number of alleles ($\overline{A}$)*

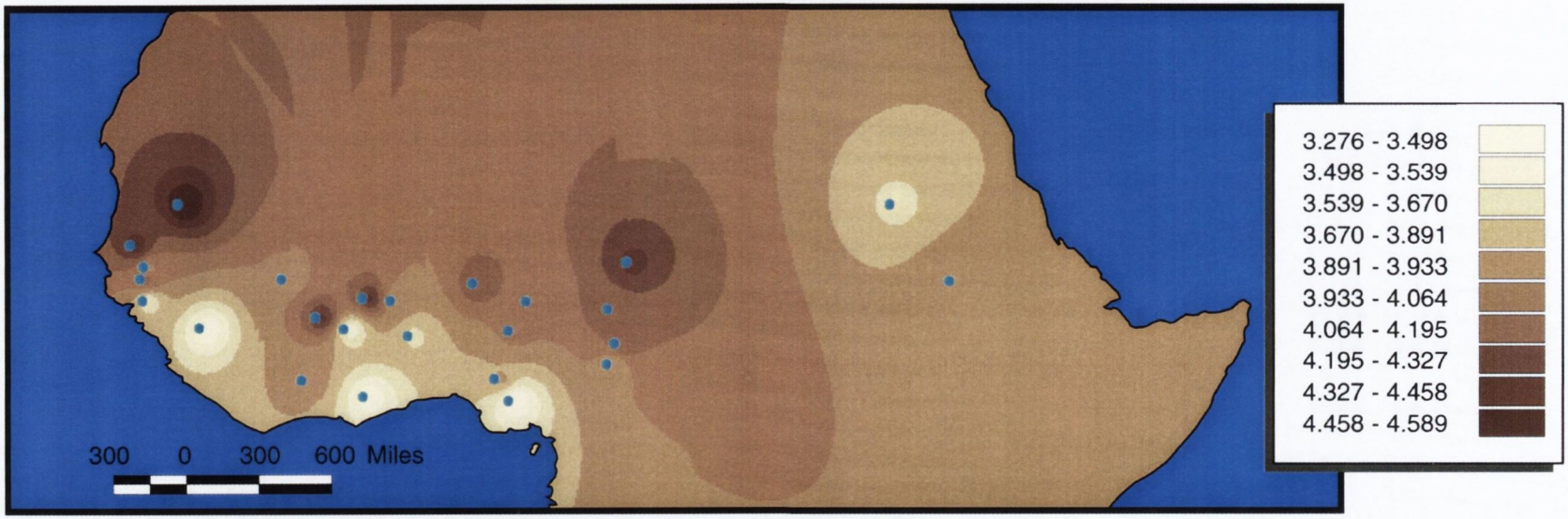


Figure 3.3 *Multi-dimensional scaling plot of F_{ST} values. Blue shading around the breed code designates breeds that are found in the Western part of West Africa. Those from the more eastern part of the region are shown by grey shading behind the breed code. Borgou, Laguna and Somba are found in central West Africa.*

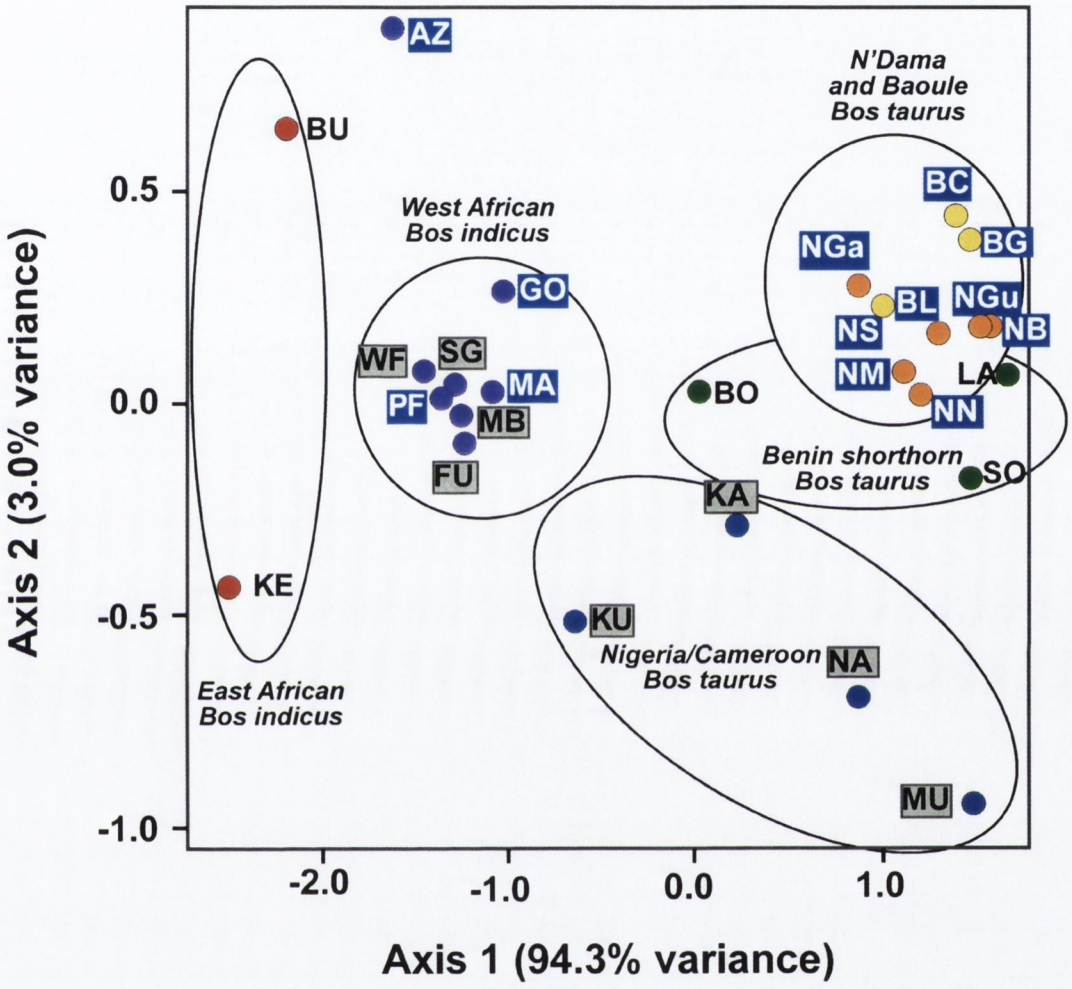
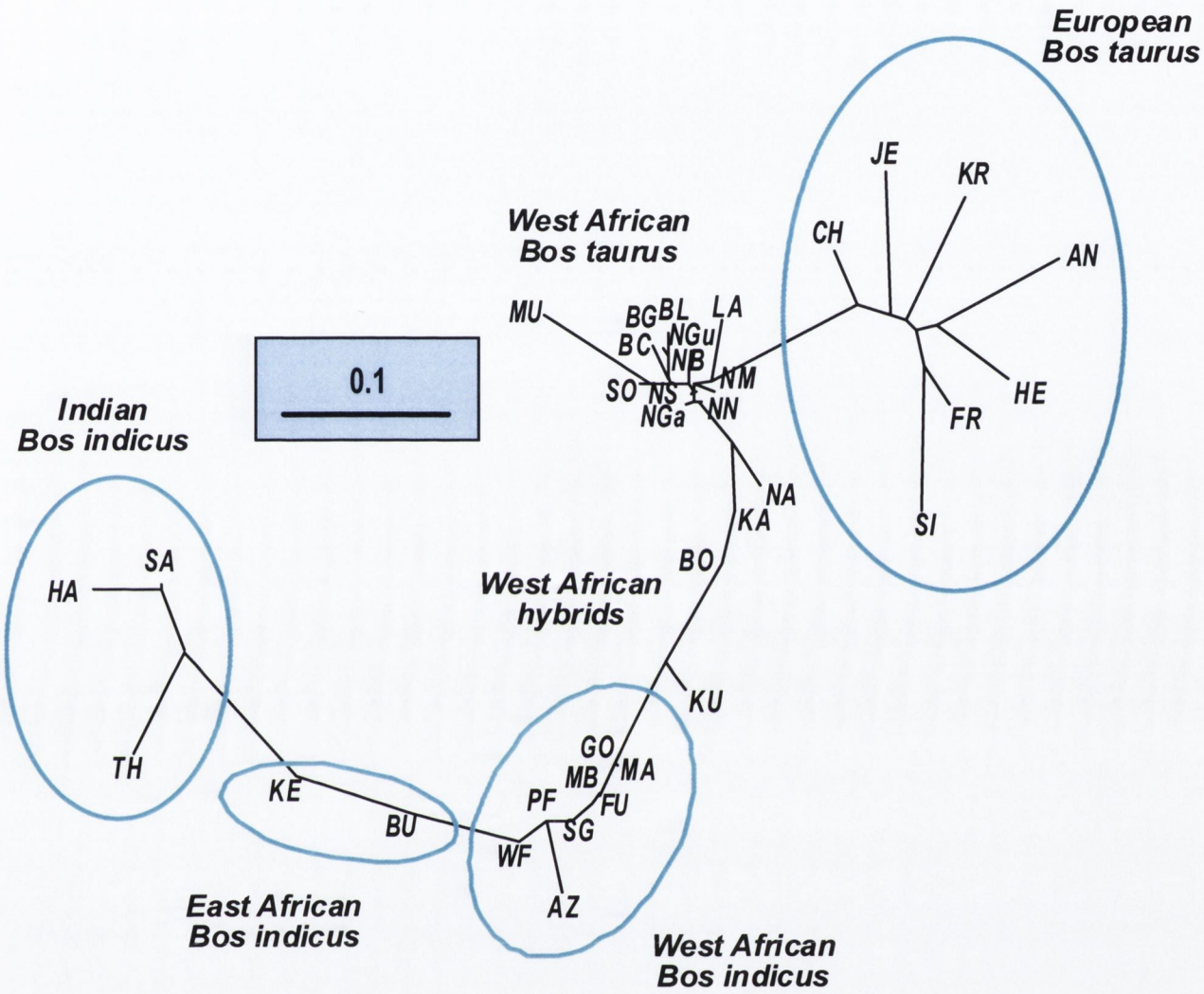


Figure 3.4 *Neighbor-joining tree constructed using a matrix of Nei's standard genetic distances between world cattle breeds. Note the truncated branch lengths in the West African hybrid group, indicative of their introgressed status (CAVALLI-SFORZA et al. 1993; MOUNTAIN et al. 1992). Geographically important groupings are indicated.*



3.4.4 ADMIXTURE

The introgression of *B. indicus* genetic material into this part of the African continent has been documented previously and is the major phylogeographic feature of the data (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). The enlarged dataset in this study enables better resolution of this effect. It also allows investigation of any possible influence from the Near East or Europe and segregation of this effect from the original background of

African *B. taurus*, the result of an independent domestication of cattle in Africa. Seven different methods were used to investigate *B. indicus* influence and two analyses to estimate any European or Near Eastern taurine influence. The remaining African *B. taurus* genetic background, which includes the trypanotolerant N'Dama breeds, is demonstrated to be one axis of the three-way admixture analysis.

3.4.4.a *Introgression of Bos indicus*

Several measures were utilised to estimate the impact of *B. indicus* introgression on the cattle breeds of West Africa. These results of these different analyses are shown in **Figure 3.5**. The correlations between the different measurements are shown in **Table 3.2**. Some of the admixture analysis methods require the assignment of parental populations while others infer most likely parental populations.

3.4.4.a.i mR

The values of *mR* for each breed are shown in **Figure 3.5a**. The values correlate highly with other admixture methods which show proportions of *B. indicus* ancestry to be very high in East African breeds and very low in the West African trypanotolerant breeds found in the tsetse zone. Some of the breeds in the study were also characterised by Hanotte and colleagues (2002). In that survey *mR* was calculated using 15 microsatellite markers, two of which are held in common with the present study. Regression analysis with the eleven breeds that were included in both studies (BG, NGu, SO, NA, KA, KU, GO, MA, MB, SG, BU) shows a highly significant relationship between *mR* values even though they have been calculated using different batteries of microsatellite markers (p=1.52 x10^8, r^2 *(adj)* = 0.97)

Table 3.2 *Table of correlations for measures of Bos indicus admixture. Twenty-six pair wise comparisons of all measures of zebu admixture were carried out and all were found to be highly significant. The correlation value has a positive value if the 2 parameters tend to move together, but a negative value if they are inversely correlated.*

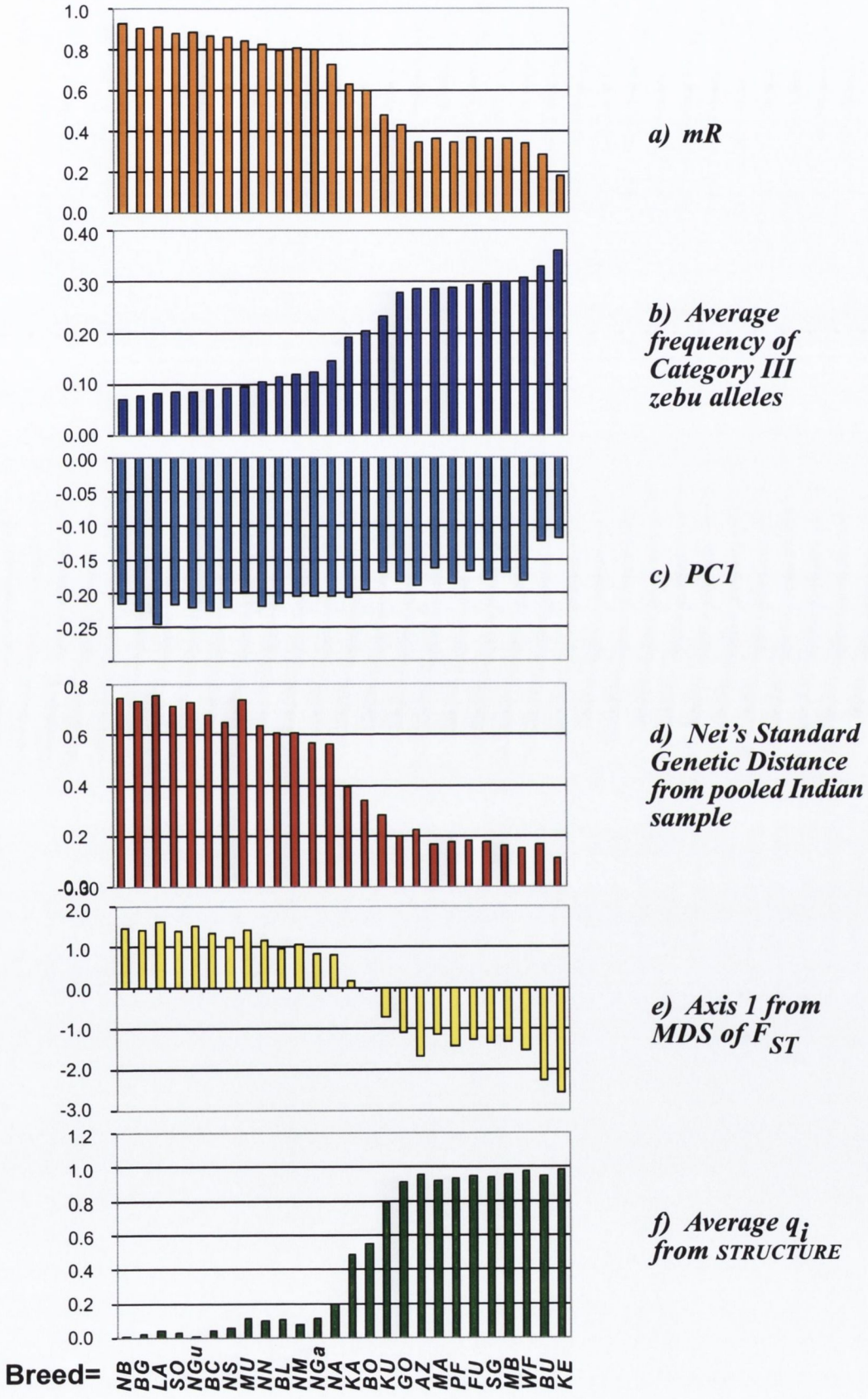
		mR	***PC1***	***Average q_i***	***D_S from Indian Bos indicus***	***Axis 1 from MDS of F_{ST}***
Average frequency category III zebu Alleles	**Pearson Correlation**	-0.998	0.877	0.986	-0.988	-0.993
	Sig. (2-tailed)	1.68×10^{29}	4.08×10^{9}	2.52×10^{20}	5.11×10^{21}	6.75×10^{24}
mR	**Pearson Correlation**		-0.878	-0.987	0.986	0.994
	Sig. (2-tailed)		3.76×10^{9}	2.11×10^{20}	3.74×10^{20}	1.06×10^{24}
PC1	**Pearson Correlation**			0.826	-0.836	-0.890
	Sig. (2-tailed)			2.10×10^{7}	1.10×10^{7}	1.11×10^{9}
Average q_i	**Pearson Correlation**				-0.987	-0.976
	Sig. (2-tailed)				2.21×10^{20}	1.68×10^{17}
D_S from Indian Bos indicus	**Pearson Correlation**					0.975
	Sig. (2-tailed)					2.85×10^{17}

3.4.4.a.ii Population-Associated Alleles

Population-associated alleles are a measure of admixture that is predominantly successful when there are large allele frequency differences between the two parental populations. *B. indicus* and *B. taurus* cattle are sufficiently divergent that there are many loci which show substantial allele frequency differences between the two sub-species. These alleles are denoted 'zebu-specific alleles' and have been used in previous admixture studies and produced results that have been supported by mitochondrial DNA and Y-chromosome studies (BRADLEY *et al.* 1994; LOFTUS *et al.* 1994b; MACHUGH *et al.* 1997). However, these analyses did not take a systematic approach and used limited parental populations.

Here alleles that are likely to be ancestry informative have been selected by picking alleles that show a frequency differential between a pooled population of European taurine cattle and the pooled data from Indian *B. indicus* breeds. Zebu alleles are defined as those that have a greater frequency in the *B. indicus* breeds than in the taurine breeds. This process aims to remove the inherent bias of selecting PAAs on the basis of high frequency in the populations under study by using populations external to the study region in the alleles selection process. PAAs have been found to be informative in other studies of admixed populations (MCKEIGUE 1998; PARRA *et al.* 2001; PARRA *et al.* 1998). The average frequency of the category III zebu PAAs is used as a measure of admixture and is shown in **Figure 3.5b**. A very high average frequency of zebu PAAs are found in the East African populations while low levels are observed in the parts of West Africa where *B. taurus* types are found.

Figure 3.5 *Histograms of six admixture measures of Bos indicus introgression in African cattle breeds.*



3.4.4.a.iii Principal Components

The first principal component (PC1) accounts for 79.6% of the variation in the data and shows the influence of Asian zebu introgression into the African continent (**Figure 3.5c**). This is directly comparable to the first principal component of Hanotte *et al.* (2002).

3.4.4.a.iv Genetic Distance

The average genetic distance from a parental population can be used as a measure of admixture as the breeds with the highest influence from *B. indicus* will have a shorter genetic distance to pure Indian zebu than those with lower levels of admixture. A pooled sample of *B. indicus* was used and the average genetic distance between the pooled sample and each of the Indian breeds was used as a measure of admixture. The results are shown in **Figure 3.5d**.

3.4.4.a.v Multi-Dimensional Scaling

Axis 1 of the MDS can be used as a measure of *B. indicus* admixture. It accounts for 94.3% of the variation in the data, demonstrating the dominant nature of this effect, and has been plotted alone in **Figure 3.5e**.

3.4.4.a.vi STRUCTURE analysis

While the methods described previously require the selection of a parental population, the Bayesian approach of the STRUCTURE program both infers parental population allele frequencies and estimates the mean of the posterior distribution of each individual's admixture coefficient ($\hat{q}$) (PRITCHARD *et al.* 2000a). Assumptions that are made in the model are outlined in **Chapter 2**. Three parental populations were assumed (k=3) for the admixture analysis after analyses with different values of k showed that the log likelihood began to plateau at k=3 (**Figure 3.6**). The authors recommend using the lowest value of k

that will satisfactorily explain the structure in the data and which appears to be biologically meaningful. The three populations inferred by the STRUCTURE analysis are characterised by African *B. taurus*, European/Near Eastern *B. taurus* and Indian *B. indicus* cattle by examining the breeds in which they have mean q values close to one. As expected, some individuals have admixture coefficient derived from two or all three inferred populations. The different coefficients are designated q_a (African taurine), q_e (European/Near Eastern taurine) and q_i (Indian zebu).

One of the axes within the data is once again shown to be that between African and Indian populations. The correlation of this measure with the other multivariate analyses indicates that it is a major trend in the data (**Table 3.2**). The coefficient of zebu admixture was denoted q_i and is 1 in pure *B. indicus* and 0 in individuals with no *B. indicus* component. This value was plotted as a contour map and is shown in **Figure 3.7**. The pattern that was observed with PC1 of MacHugh *et al.* (1997) and Hanotte and colleagues (2002) is also evident here. q_i is one of the measures of zebu admixture that are shown in **Table 3.2**. It is reassuring to see that all of these measures are highly correlated ($p<0.001$). These measures provide independent tests for the amount of influence of *B. indicus* in Africa. Our choice of parental populations for PAAs and *mR* is supported by the correlation between these measures, and the integrated Bayesian approach carried out by the STRUCTURE program. As has been shown previously (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997), the highest indicine influence is in the east of the continent and the values of q_i are lowest in the trypanotolerant breeds from the coastal countries of West Africa, where the tsetse fly is endemic. There is a gradual cline of zebu influence between these two extremities. In West Africa there is a gradient from north to south as the native African taurine continue to predominate in the tsetse infected coastal regions (**Figure 3.7**).

Figure 3.6 *Line plot of log likelihoods of different values of k (numbers of parental populations) for* STRUCTURE *simulations. For both runs, the values start to asymptote at K=3, so this is likely to be the lowest value of k that satisfactorily explains the structure in the data.*

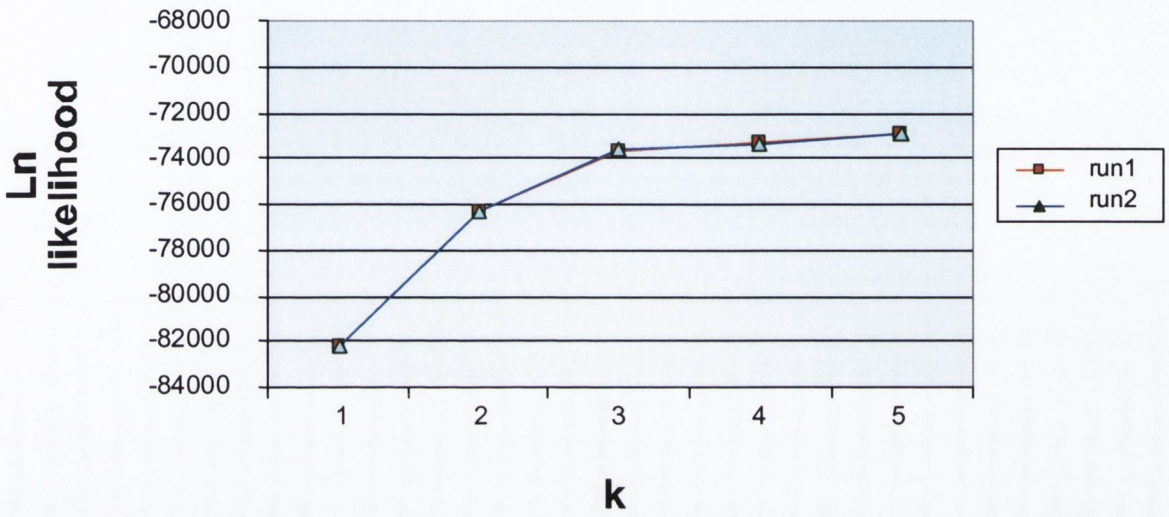
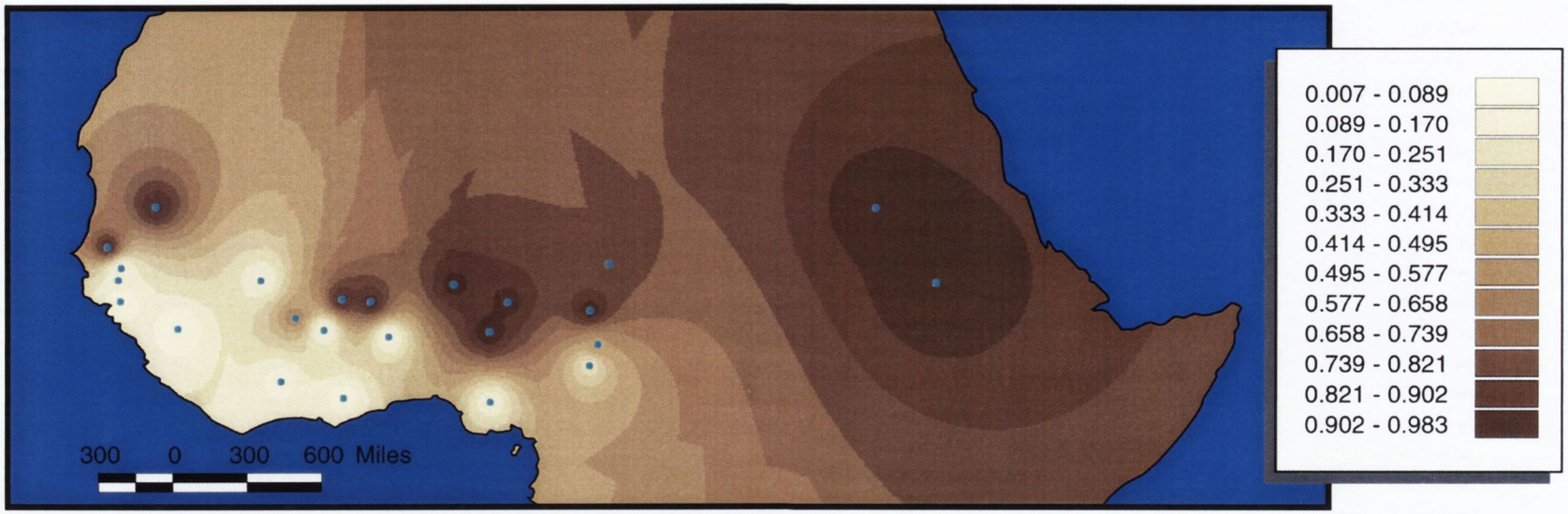


Figure 3.7 *Contour map of mean q_i, the admixture coefficient characterised by Indian Bos indicus.*



3.4.4.b *Introgression of European/Near Eastern B. taurus*

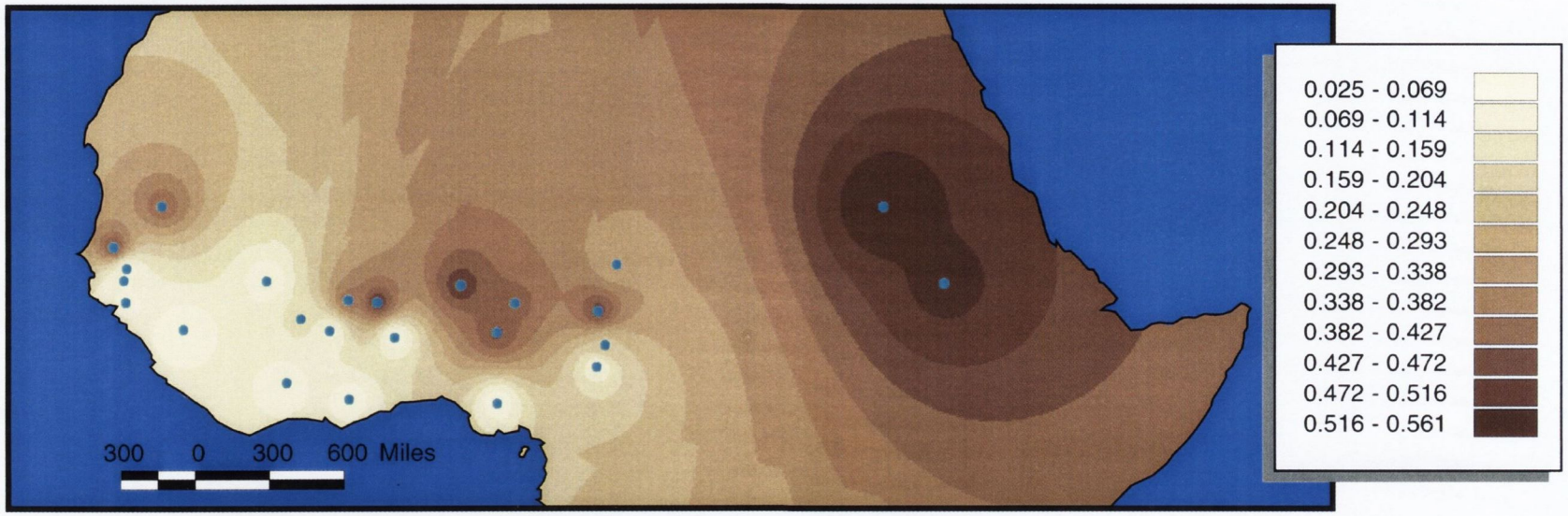
The majority of studies carried out on the cattle of this region focus on the very strong bi-directional clines between *B. indicus* and *B. taurus* (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). However, there is evidence for African influence in Iberian cattle breeds (CYMBRON *et al.* 1999) and it is possible that there may have been reciprocal transfer of genes from Europe to Africa. However, the majority of the exogenous post-domestication taurine influence in Africa is likely to have come via the Near East. It represents a major challenge to dissect any alternative taurine introgression from the underlying substrate of indigenous African taurine ancestry that was originally the sole variety on the continent. Extensive *B. indicus* introgression into the Near East has also occurred (LOFTUS *et al.* 1999), so it is reasonable to compare African taurine with European taurine. European *B. taurus* breeds represent a sub-set of the diversity that would have been present in the pure taurine populations of the Near East (TROY *et al.* 2001). The European/Near Eastern taurine influence was explored by using the multivariate principal component analysis and the Bayesian approach as implemented using the STRUCTURE algorithm.

3.4.4.b.i STRUCTURE analysis

Another axis of the three-way Bayesian analysis that is of interest is that between European and African taurine breeds. This should account for any European or Near Eastern introgression into the African continent. The level of European or Near Eastern influence was measured by q_e, which is 1 in pure European *B. taurus*. The value of q_e was normalised by expressing it as a proportion of the total taurine admixture level. The total taurine admixture coefficient is equal to $q_e + q_a$, the sum of the European and African elements. This will be very small in predominantly *B. indicus* breeds. The normalised q_e (q_e/N) allows us to estimate what proportion of the taurine component is European/Near Eastern and is shown as a contour map in **Figure 3.8**. The absolute value of q_e is a

measure of total European or Near Eastern taurine input, but does not take account of either q_i or q_a. The pattern of q_eN parallels the distribution of q_i. It is high in East African *B. indicus* breeds, and in hybrid breeds sampled around the boundaries of the West African tsetse zone. As with the zebu component, the levels of normalised q_eN are low in breeds found in the southwestern coastal countries of Africa.

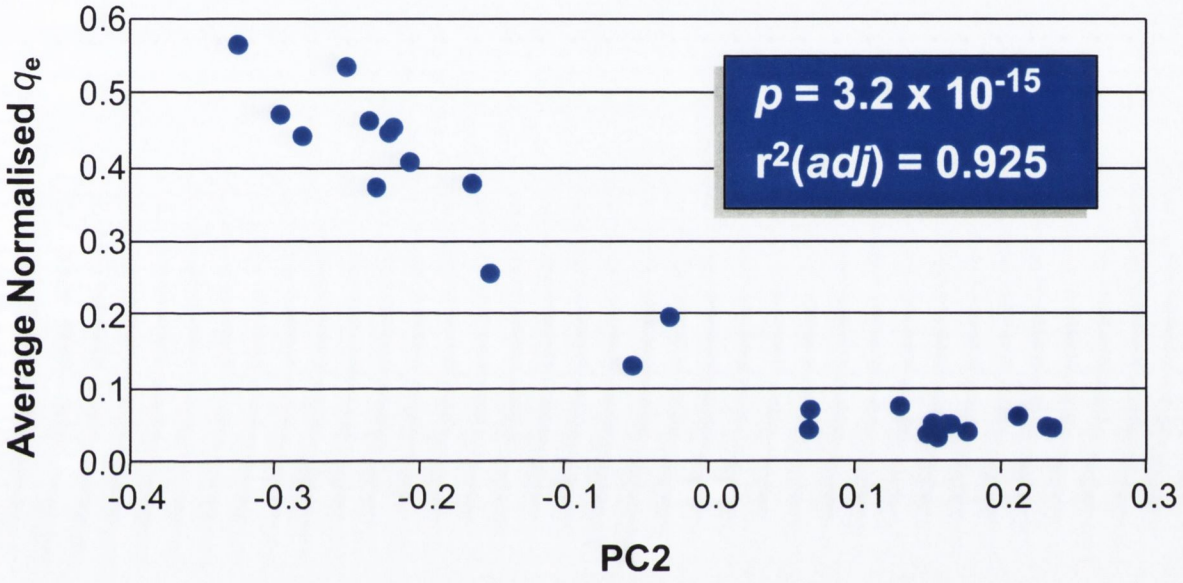
Figure 3.8 *Contour map of a normalised value of q_e the admixture coefficient characterised by European/Near Eastern Bos taurus.*



3.4.4.b.ii Principal Components.

A regression analysis of q_eN and the second principal component (PC2) (p=3.24 x 10^{-15}, r^2(adj)=0.925) supports the hypothesis that PC2 represents the influence of Europe and the Near East (**Figure 3.9**).

Figure 3.9 *Regression scatter plot of the second principal component and the normalised values of* q_e.



3.4.4.c African B. taurus – the genetic background.

Previous studies have indicated that the original cattle of the African continent were probably of a *B. taurus* type and were derived from an independent domestication of local aurochsen. The present day pattern of African-type *B. taurus* is illustrated by the distribution of q_a and is shown as a contour map in **Figure 3.10**. The map clearly shows the opposite pattern to q_i (**Figure 3.7**) and clearly demonstrates how the continuing and widespread movement of zebu cattle from the east is diminishing the native African gene pool.

Figure 3.10 *Contour map of q_a, the admixture coefficient characterised by African Bos taurus*

3.5 DISCUSSION

The results presented in this chapter complement previous genetic analysis of the cattle of West Africa. As previous authors have discovered, there is a cline of *B. indicus* introgression in this area, which has been created by the westward migration of zebu cattle hybridising with indigenous taurine cattle (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). The increased numbers of populations in the current study produces a more detailed picture of diversity in West Africa. While it is likely that cattle were domesticated on the African continent (WENDORF and SCHILD 1994), there are several complicating factors in definitively assigning archaeological material as being from either domestic or wild animals (CLUTTON-BROCK 1989; EPSTEIN 1971). These include sexual dimorphism and the close relationship that would have existed between the earliest domesticates and their wild relatives (BÖKÖYNI 1989). In the light of this, genetic evidence is a useful tool to confirm whether cattle in Africa have had genetic contribution from local domestication, and possibly to elucidate the patterns of any past introgression events.

3.5.1 GENETIC DIVERSITY AND ADMIXTURE

In the absence of subsequent input, an older source population will typically display greater genetic diversity than a population derived more recently from it (JORDE *et al.* 1998). Selection may also act to reduce diversity, while hybridisation will increase diversity by bringing together alleles from two distinct lineages. A sample size-corrected mean number of alleles ($\overline{A}$) was calculated in the African, European and Indian samples (**Table 3.1**). As expected, hybrid breeds with both a substantial *B. taurus* and *B. indicus* component have the highest $\overline{A}$ values (**Figure 3.2**). More unexpectedly, the East African indicine breeds (Butana and Kenana) have comparable diversity levels to the purer West African taurine breeds.

There is a belt of high diversity levels around the perimeter of the tsetse fly infested zone, which implies that there may be continual introgression occurring in this vicinity. The value of $\overline{A}$ in European taurine breeds is comparable to some of the non-hybrid African taurine breeds, but lower than most. The lowest diversity values are seen in the endangered Kerry breed from Ireland and the island-bound Jersey breed (**Table 3.1**). For the most part, however, the present results show that of the previously mentioned factors that influence diversity, as measured by $\overline{A}$, admixture is the predominant one in Africa.

Hanotte and colleagues (2002) found that three orthogonal components of variance accounted for the majority of the variation in their data from a continent-wide sample of African cattle. They interpreted these as the signature of local domestication in Africa followed by separate genetic input from centres of domestication in the Near East and India. In the current analysis, the diversity in Africa was also classified into components and the results correspond.

Our findings agree with those of previous authors, who found that the indicine/taurine divergence is the foremost trend in African autosomal microsatellite data (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). The very high levels of *B. indicus* admixture seen in breeds such as Borgou, Namchi, Gobra and Azawak, which are found around the perimeter of the tsetse infested zone, are consistent with major incursion of zebu into this region. The lack of indicine introgression into breeds that are found within the tsetse belt is a result of the inability of *B. indicus* cattle to survive in areas where typanosomiasis is endemic. The hybrid individuals have higher levels of linkage disequilibrium (LD) at both linked and non-linked microsatellite markers (see **Chapter 4**), and higher levels of diversity as

described in the previous section. This LD is utilised by the STRUCTURE method to detect likely hybrid and parental populations.

It is a commonly held view that *B. indicus* first entered the African continent via the Suez route from Arabia (EPSTEIN and MASON 1984). Our results are consistent with those of Hanotte and colleagues (2002) who suggested that this initial introduction had little impact on West Africa. Some authors believe that the initial number of humped cattle that were introduced was quite low and that they were rapidly absorbed into an indigenous taurine gene pool, resulting in the production of sanga breeds (PAYNE and WILSON 1999). More recent research has demonstrated that the major incursion of *B. indicus* to Africa was mediated via Arabia and maritime routes to the Horn of Africa (HANOTTE *et al.* 2002). There is a substantial time gap, possibly over 1000 years, between the initial introduction of zebu to the continent and their arrival in the western portion. This means that the observable *B. indicus* introgression in West Africa may transpire to be the result of migration from East to West of pastoralists (such as the Fulani) conveying hybrid cattle (CLUTTON-BROCK 1989; HANOTTE *et al.* 2002). Previous work has found support for two differing patterns of *B. indicus* introgression, which may be associated with different rates of migration through the continent (HANOTTE *et al.* 2002).

MDS (**Figure 3.3**) analysis of the numerous breeds that are included in the current study shows genuine clustering, that is related to both breed geography and history. The low level of *B. indicus* admixture in the trypanotolerant N'Dama populations is almost certainly the result of selection against introgressing breeds in the humid tsetse regions of West Africa. Our results do not shed much light on the putative origins of N'Dama cattle, but suggest that they represent the extremity of a continuum of indigenous African taurine

to which most of the shorthorn breeds also belong. MDS groups the N'Dama breeds from different countries in a cluster, which implies that few genetic modifications have occurred as the breed spread throughout the region. The adjacent breeds include Baoule, Laguna, Borgou and Somba, all shorthorn taurine breeds. The second axis of variance in the MDS, which seems to separate breeds in the east of the study region from those in the west, could be related to q_e, the influence of *B. taurus* from the Near East, which has migrated to West Africa, alongside the indicine component, in hybrid individuals.

Bayesian analysis seems to support the division of genetic variance in West African cattle into three components (**Figure 3.6**). It may be argued that this output a) reflects the ancestral divergence between *B. taurus* and *B. indicus* and b) also lends support to the evidence for an independent African domestication event. Whereas most studies of cattle in this region exploit the ancient bifurcation between *B. indicus* and *B. taurus* cattle to differentiate and describe breeds, it is far more challenging to analyse low-level admixture between more closely related lineages, such as the African and European/Near Eastern *B. taurus* lineages. Our second principal component appears to be the same as the third principal component of Hanotte et al. (2002). The larger sample in that paper included a wider sampling from the entire African continent and PC2 was interpreted as the African taurine component of variation. This element of variation stretched from the northwest to the south of the continent, and is detected in our data by the q_a axis of the Bayesian analysis (**Figure 3.9**).

A normalised version of q_e was used (**Figure 3.8**) to describe the location of any *B. taurus* influence derived from Europe or the Near East. This admixture coefficient is coincident with q_i, the *B. indicus* genome component. It seems likely from a comparison of **Figures**

3.7 and 3.8 that the pattern of the introgressing *B. taurus* can be explained by taurine genes from the Near East 'hitchhiking' into West Africa, after incorporation into hybrid breeds of East Africa. The highest absolute value of q_e (not scaled according to the African taurine element) was recorded for the West African breeds in the Cameroon, Guinea Bissau, Ghana, The Gambia and Senegal and is approximately 5%. This present study underlines the need for additional sampling from Northern African countries such as Libya and Algeria, as these would contribute greatly to gaining a clearer picture of taurine gene flow in North and West Africa, and between Africa and Europe.

3.5.2 CORRESPONDENCE OF ADMIXTURE MEASURES.

There has been much debate about the most appropriate methods to measure admixture, mainly of which are confounded by lack of information about contributing parental populations (BIERNE *et al.* 2003). In order to estimate the amount of input from an ancestral population it is necessary either to assume that an extant population is a suitable model of the ancestor, or to infer the probable parental allele frequencies. In order to get confident estimates about admixture proportions it is desirable to have independent measures that correlate. Here six methods to estimate *Bos indicus* admixture proportions in West African cattle breeds were used. These are listed in **Table 3.2**, which includes correlation values for each pair of measures. All of the measures are highly significantly related.

A considerable obstacle to any admixture study approach is selection of the parental population. The selection of private alleles could be biased by the incorrect assignment of parental or hybrid status. The difficulties in estimating population and especially individual admixture when gene frequencies in the parental population before the admixture event are unknown have been discussed previously (ESTOUP *et al.* 1999). Pooled Indian and pooled

European samples were used as parental populations for the ADMIX and PAA analysis. The use of these representative parentals can be justified by previous findings (BRADLEY *et al.* 1996; HANOTTE *et al.* 2002; MACHUGH *et al.* 1997). However, more importantly the breeds under study are not used in the process of selecting the PAAs. In previous studies (MACHUGH *et al.* 1997), 'zebu-diagnostic' alleles were defined by observation of allele frequencies in both putative parentals and putative hybrids thereby creating an inherent bias. They were defined as having high frequency in Indian and East African zebu, low frequency in African taurine and absent or marginal in European cattle. Using the present method, the average frequency of PAAs in the hybrid breeds represents a more satisfactory and more impartial measure of admixture.

Nei's standard genetic distance from the Indian *B. indicus* sample and F_{ST} values between the African breeds are further independent measures of the relationships between the African breeds, as each is compared individually to every other breed in the two analyses. For the former, the Indian *B. indicus* acts as an outgroup to quantify the amount of zebu introgression into the African breeds by measuring the genetic distance between each of them and each Indian breed and averaging this distance. The multivariate analysis methods, MDS of genetic distance and principal components, do not utilise any genealogical or parental information but, in designating axes that account for variation in the data, can determine the major trends and groupings. The Bayesian approach taken by STRUCTURE does not require pre-assignment of parental populations, and indeed estimates the likely ancestral allele frequency distributions of the parentals. Its excellent correspondence with the other measures suggests that these other methods are accurate if selection of contributing populations and outgroups are correct.

The correspondence between admixture measures as applied to the other axes of variance is not as good as for the major trend in the data. Our results imply that when an admixture effect is large, the method of detection is not so critical, but that when a contribution is smaller, some methods are superior. It is far more difficult to accurately estimate the impact of European admixture as the extent is much smaller, and different admixture estimation methods produce results that are not highly correlated. These estimates tend to correlate well when the level of admixture is very high or very low but do not match well when admixture levels are intermediate.

3.5.3 APPLICATION OF ADMIXTURE STUDIES TO WEST AFRICA

Reliable estimation of admixture levels has been the focus of much interest recently as it may prove extremely useful in the detection of genes involved in disease susceptibility and resistance (CHAKRABORTY and WEISS 1988; MCKEIGUE 1997; PARRA *et al.* 1998; SHRIVER *et al.* 1997). A sub-structured hybrid population, which has had a genetic contribution from one parental population with a particular disease phenotype, may be used to detect the genes involved in that phenotype, as they will be co-transmitted with other markers from that parental population. High allele frequency differentials between two contributing parental populations at neutral markers can be used to try and detect genome regions that could be involved in the phenotype of interest. This approach to map genes is most suitable when the two contributing parental populations have both high neutral allele frequency differential and high disease frequency differential.

Trypanosomiasis is a devastating disease which is responsible for the deaths of millions of livestock animals each year and a reduction in the productivity of many more (see **Chapter 1**). Each year more than US$1 billion dollars are spent trying to control the disease and in direct production losses (FAO 2003b). The trypanotolerant trait that is present in

indigenous West African taurine cattle is not present in the introgressing *B. indicus* cattle and is dependent on admixture proportion in hybrid breeds. This indicates that admixture-generated linkage disequilibrium might be an appropriate tool to try and locate genes that contribute to trypanotolerance. The initial phase of any mapping project would be to obtain accurate admixture estimates for properly defined, biologically distinct groups. It is not inconceivable that detection of the genes that enable African taurine cattle to withstand the trypanosome infection could also contribute to therapies for human use (KEMP 2002). Current estimates suggest that sleeping sickness threatens around 60 million rural Africans (FAO 2003b). It is evident that the native taurine substrate of domesticated cattle has almost completely disappeared through diseases, such as rinderpest, or absorption into the *B. indicus* gene pool. Consequently, studies such as this reinforce the need to protect the remaining non-introgressed West African breeds from hybridisation as they represent an important genetic resource in efforts to combat trypanosomiasis.

CHAPTER 4

HAPLOTYPE BREAKDOWN IN THE BOVINE HYBRID ZONES OF AFRICA AND THE NEAR EAST: linkage disequilibrium and applications to admixture mapping

4.1 INTRODUCTION

Since the initial domestication of cattle, substantial genetic exchange has taken place between the original populations, resulting in the production of hybrid zones between the two primary lineages: *Bos taurus* and *Bos indicus*. The current study focuses on several groups of previously studied breeds with known hybrid backgrounds, which are compared to non-introgressed populations from the taurine and indicine lineages.

Admixture, as the result of introgression events, has created hybrid zones in the Near East, and in over large areas of the African continent (HANOTTE *et al.* 2002; LOFTUS *et al.* 1999; MACHUGH *et al.* 1997; MEGHEN *et al.* 1994). In general, the introgression has been classed as *B. indicus* incursion into a static *B. taurus* background, but in reality the situation is likely to be more complex, with some reciprocal gene flow and movement of populations. However, for simplicity it is useful to initially consider the present hybrid populations as the former model. European and Indian populations have escaped significant secondary introductions and are principally included in the present survey as control samples. The histories of the hybrid zones in West Africa and the Near East have been detailed elsewhere (LOFTUS *et al.* 1994a; MACHUGH *et al.* 1997; MEGHEN *et al.* 1994), and in previous chapters of this thesis. Eastern and southern African breeds are also included in the present analysis so the history of these populations is discussed below.

4.1.1 CATTLE IN EASTERN AND SOUTHERN AFRICA

It has been previously demonstrated (HANOTTE *et al.* 2002) that the cattle of southern and western African share a common indigenous African genetic component (see also **Chapter 5**). However, while archaeological evidence reveals that cattle first arrived in West Africa around 4000 BP (MACDONALD and HUTTON MACDONALD 2000), the earliest artefacts that

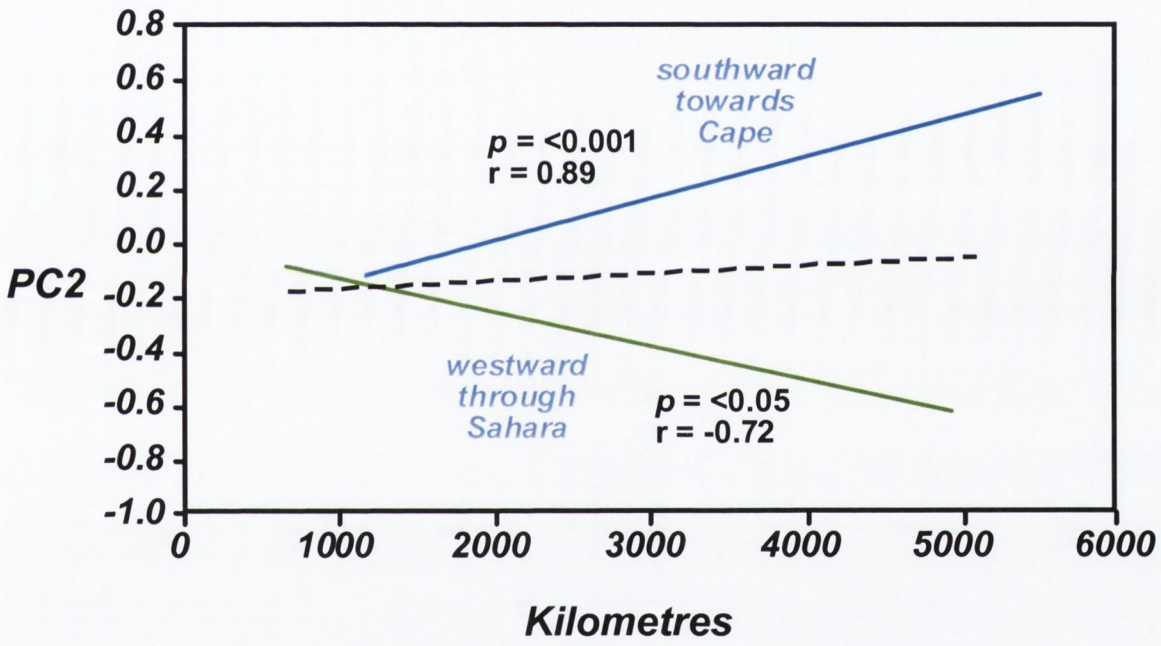
document the arrival of cattle in southern Africa date from only 1600 BP (CLUTTON-BROCK 2000; SMITH 2000). In fact, it is likely that southern Africa was the last part of the continent to acquire cattle pastoralism (SMITH 2000). The original cattle in both of these regions were humpless (MACDONALD and HUTTON MACDONALD 2000; SMITH 2000) and probably the descendants of *B. taurus* cattle that were domesticated in the north-eastern Sahara before 8000 BP (MARSHALL and HILDEBRAND 2002). Disease is thought to have been a considerable obstacle to the entry of domestic cattle to southern parts of Africa, which were only established in the area in sizeable numbers with the rapid spread of Early Iron Age farmers (GIFFORD-GONZALEZ 2000; MARSHALL and HILDEBRAND 2002).

However, it is still uncertain whether the first domestic cattle were those brought from the Cameroons by these Iron Age people, or by East African herders who moved south along an eastern route. Stylistic traits associated with non-Iron Age East African pottery have been associated with some of the earliest pastoralists in the Cape, but Iron Age remains seem to have masked most traces of a previous herding culture (SMITH 2000). Bantu speakers were first established around the Lakes area of East Africa less than 2000 BP (RICHARDS *et al.* 2003). Most authors agree that herding spread southward from the area that is now Uganda from this time onwards, and was associated with Bantu culture (HANOTTE *et al.* 2002).

Hanotte and colleagues (2002) have suggested that there were different modes of dispersal of taurine cattle from their hypothesised domestication centre in the northern Sahara, using evidence from multivariate analysis. For the second principal component, they found distinct significant correlations between taurine breeds that are located either westward or southwards from a putative centre of domestication (**Figure 4.1**). Migration from the

northeast, including subjection to diverse disease threats, is considered to be responsible for the present differentiation between these two groups of African *B. taurus*.

Figure 4.1 *(reproduced from* HANOTTE *et al. 2002) Regression of principal component 2 (PC2), versus distance from putative origin of African pastoralism (calculated by the distance of each breed from the most north-easterly African population sampled). PC2 was found account for 9.5% of total allelic variation and is hypothesized to be the major axis of African B. taurus variation. The dashed line represents all zebu breeds, which do not show a relationship to PC2. However, there are two significant correlations; these represent two dispersals of indigenous African taurine cattle from a domestication centre in the North of the continent. One shows B. taurus breeds that are between the putative origin and West Africa (green line), while the other represents breeds that are found along a route that stretches to the South (blue line).*



The cattle that are found around the Horn of Africa have *B. indicus* type morphology and derive a large portion of their genome from imported zebu. There is scant reliable archaeological evidence for the initial arrival of humped cattle in east Africa. This is because most relevant discoveries have been of bones that are heavily processed and fragmented (MARSHALL 2000). However, the distribution pattern of other sites and ancient texts suggest that a first wave of zebu introductions may have been migrations, while

secondary were the result of trade (CLUTTON-BROCK 1989; MARSHALL 2000). Genetic evidence supports the thesis that the major point of *B. indicus* entry was into eastern African and Madagascar, from the south of the Arabian peninsula (present day Yemen), and maritime routes, and not via the Suez (HANOTTE *et al.* 2002).

4.1.2 HYBRID POPULATIONS AND HAPLOTYPE BREAKDOWN

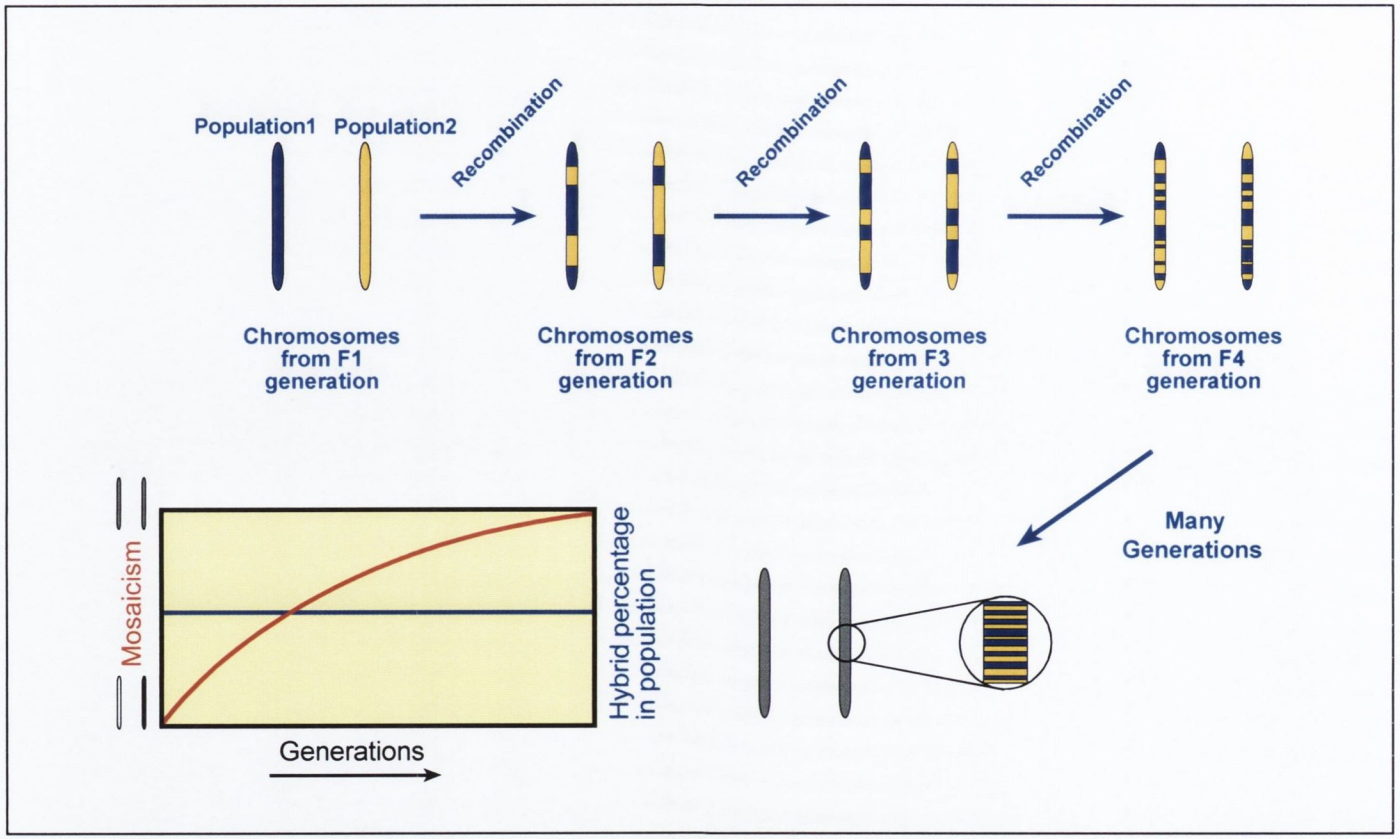
The genetic exchange between taurine and indicine cattle represents an appropriate model to study the effects of haplotypic decay after an admixture event. More importantly, studying the breakdown of haplotypes may increase understanding of how the trypanotolerant trait of West African *B. taurus* cattle is being broken down by introgression by zebu stocks. Moreover, this kind of data may be useful in assessing populations for the promising technique of mapping other genes using admixture-generated linkage disequilibrium (LD). Furthermore, the information should provide additional information about when, and to what degree, admixture has occurred within each of the studied regions.

In recent times there has been considerable interest in hybrid populations, which receive a measurable proportion of their genomes from differentiated parental populations. Therefore, over any short interval, a haplotype from an individual of mixed descent will consist mainly of alleles that are more common in one of the founder populations. However, this effect will gradually be eroded through recombination and mutation. It is likely that the populations under study have quite distinct histories of admixture. Admixture may have occurred to different degrees and at different times. The genetic integrity of the contributing populations will have a profound effect on the genome of a hybrid. Detecting the pattern of admixture is likely to be considerably more challenging if parental populations are themselves admixed. Closely linked markers can be used to

examine how parental haplotypes have been broken down over time in the hybrid population. An illustration of this is provided in **Figure 4.2**. When admixture first occurs the population will contain intact chromosomes from each of the parental populations. Over time recombination will cause these to mix and gradually become more mosaic, even though the proportion of admixture remains the same in the hybrid population.

It was suggested by McKeigue (1997) that closely linked groups of microsatellite markers spread over the genome in a hybrid population could be used to detect the pattern of admixture and that this information could be used for gene mapping studies. Completely linked microsatellites can be informative about these aspects of genealogical history (WILSON and BALDING 1998). A marker density of every 10 cM or less was estimated but the density of markers required would depend on the age of admixture (MCKEIGUE 1997). If admixture is more ancient reciprocal recombination, genetic drift and mutation will have eroded the association between markers from parental populations and a greater density of markers will be required to detect significant associations. When a hybrid population is formed, the chromosomes will gradually become more 'mosaic' as the parental haplotypes recombine and are broken down (CHAPMAN and THOMPSON 2002).

Figure 4.2 *Schematic showing haplotype breakdown over time. In a new hybrid population there will be intact chromosomes that have been contributed from each of the parental populations. At the prophase of meiosis, recombination will occur producing F1 generation chromosomes that are comprised of blocks of genetic material from each parent. As this process is repeated the size of inherited blocks will decrease until a stable hybrid is eventually formed after many generations*



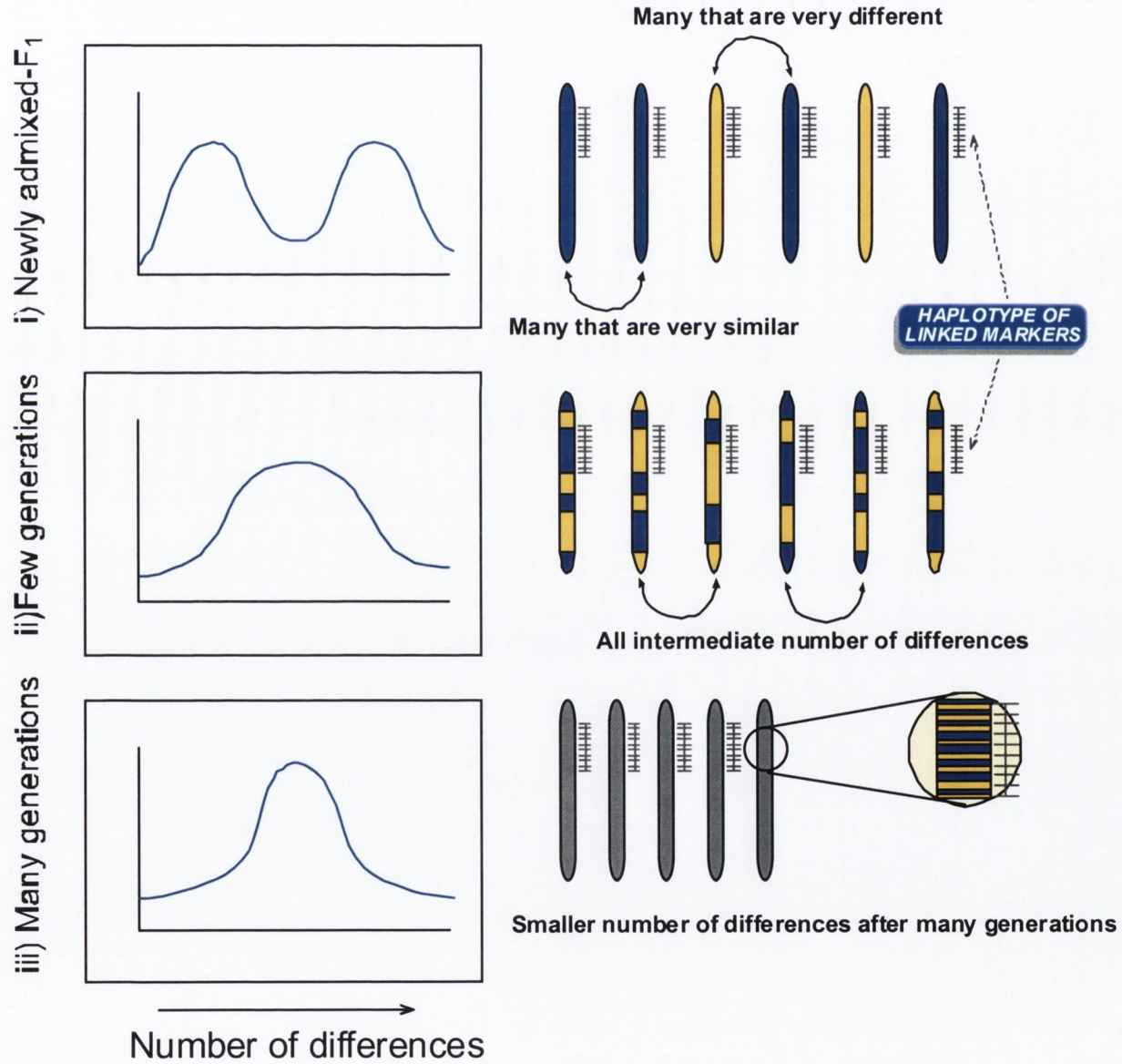
4.1.3 ADMIXTURE DETECTION

Methods to detect and determine the age of admixture include mismatch analysis, linkage disequilibrium (LD) testing and assigning probable ancestry to microsatellite alleles.

4.1.3.a Mismatch analysis

Mismatch analysis has traditionally been employed for DNA sequence analysis and examines the distribution of the observed numbers of differences between pairs of haplotypes (ROGERS 1995; ROGERS and HARPENDING 1992; SCHNEIDER *et al.* 2000). A putative model for the manner in which a mismatch distribution might evolve in a hybrid situation is outlined in **Figure 4.3**. A newly admixed F1 population will have chromosomes from both parental populations that may be very different from each other. Over time, recombination and drift will cause homogenisation of the chromosome population but the mean of the mismatch distribution may still be quite high. As the population ages and approaches a stable hybrid the variance of the mismatch curve should gradually decrease, as the alleles from the two parental populations are dispersed throughout the population. Furthermore, in the absence of subsequent introgression the number of differences between individuals may decrease. Alleles may be lost from the hybrid population, especially under a scenario where the effective population size decreases. A zero point will never be reached due to mutation and the original variation in both populations.

Figure 4.3 *Hypothetical mismatch distributions of X-chromosome haplotypes following hybridisation events. Mismatch distributions plot the frequency of the number of pairwise differences between haplotypes. Part i shows a hypothetical mismatch distribution for a newly admixed population, which contains intact chromosomes from the two founder populations. The haplotype is constructed from a group of linked markers, as illustrated below, and represents a sample from the chromosome. The left-hand peak in the bimodal distribution represents the differences between haplotypes derived directly from one parental population being compared to other haplotypes from that population. The right-hand peak is the result of comparisons between haplotypes from the two different parental populations. As recombination gradually breaks down the intact haplotypes (ii and iii), the range of numbers of differences between individuals is gradually reduced as the alleles from the original founders are dispersed through the hybrid population. In the absence of subsequent introgression, the mean number of differences between individuals may decrease.*



4.1.3.b Linkage disequilibrium

Admixture creates LD, which decays over time (CHAKRABORTY and SMOUSE 1988). The level of LD at neutral loci therefore yields indirect estimation of the hybridisation, provided a current state of neutral introgression in the populations (BIERNE *et al.* 2003). It might be expected therefore, that the different cattle hybrid populations will have measurable differences in patterns and amounts of LD as a result of their different genetic histories. If chromosomes in hybrids have been broken into many small pieces, relative to the original contributing parental populations, LD due to co-descent of markers on an intact segment will stretch over only short distances. Recent or continuous admixture will results in longer chromosomal segments that can be traced directly and unbroken to a founder parent, and high levels of LD. Previous studies in domestic cattle have revealed reasonable elevated levels of LD, although these studies have included only European breeds.

4.1.3.c Allele ancestry

As explained in **Chapter 1**, previous studies have revealed the genetic divergence between *B. indicus* and *B. taurus* to date from well before domestication and as such, there are many population associated alleles. In previous chapters, populations associated alleles have been employed to detect and quantify admixture levels. The alleles are detected based on a frequency differential between putative parental populations and their average frequency is used as a measure of admixture. Assigning alleles along a chromosome segment requires a slightly different approach, and the current study employs a Bayesian and partially Bayesian approach.

4.2 MATERIALS AND METHODS

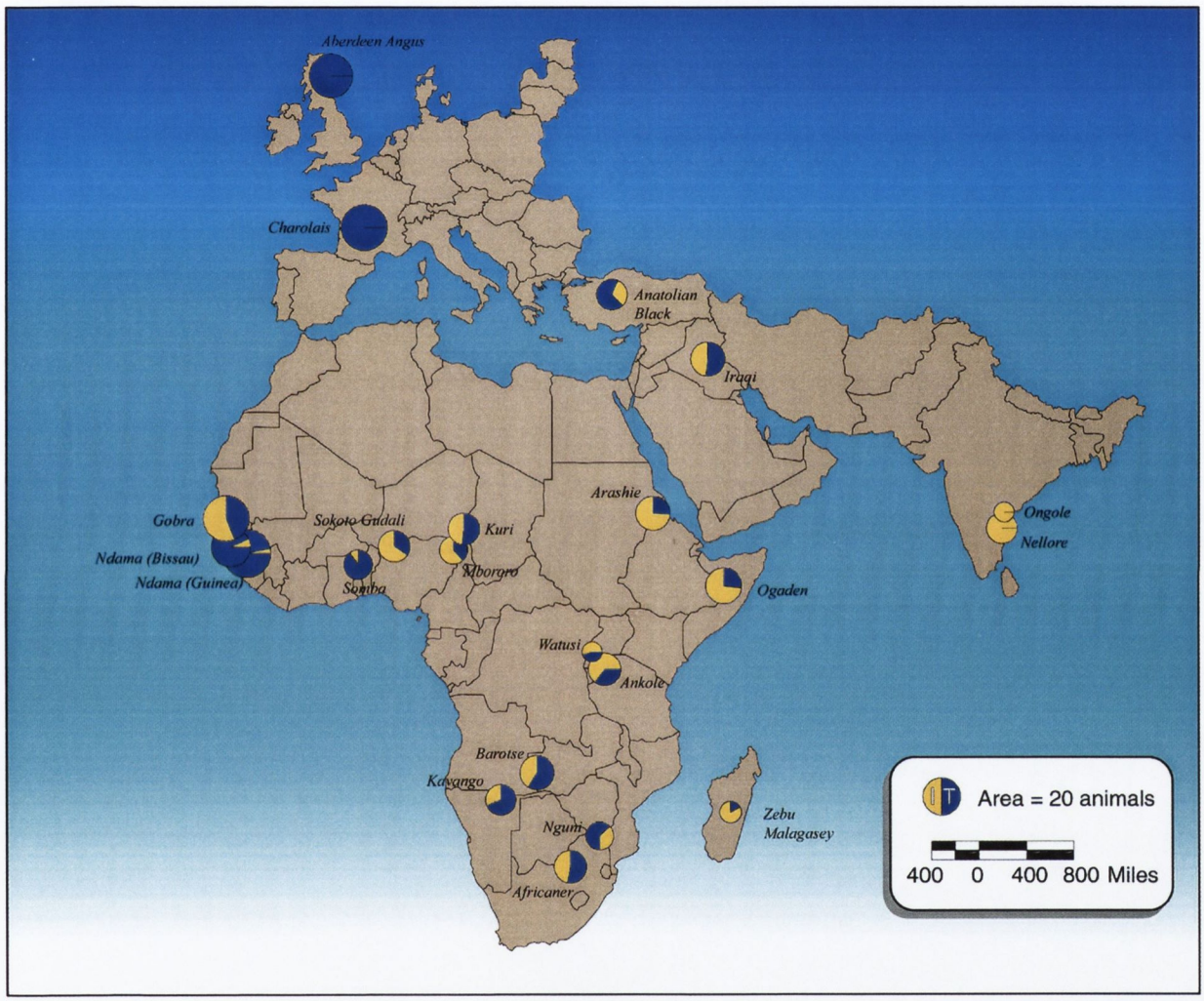
4.2.1 SAMPLE INFORMATION

346 male cattle DNA samples from 22 breeds were obtained and are listed in **Table 4.1**. These 22 breeds were pooled into nine population groups. DNA had been extracted from blood or hair samples as previously described by MACHUGH et al. 1997, LOFTUS et al. 1999 and HANOTTE et al. 2002. The approximate location of the sample sites and the previously published admixture coefficients are shown in **Figure 4.4** (HANOTTE *et al.* 2002; LOFTUS *et al.* 1999; MACHUGH *et al.* 1997; MEGHEN *et al.* 1994). For certain analyses, the novel data was combined with microsatellite data from previously published studies (HANOTTE *et al.* 2002; LOFTUS *et al.* 1999; MACHUGH *et al.* 1997; MEGHEN *et al.* 1994).

Table 4.1 *Cattle breeds that are included in the present study. N is the number of individuals in each sample. Breed that are highlighted with an asterisk were genotyped for fifteen microsatellite markers and 3 single nucleotide polymorphisms. The remaining breeds were genotyped for a subset of ten microsatellites.*

Breed name	N	Code	Regional designation
Aberdeen Angus*	29	AN	European *Bos taurus*
Charolais*	33	CH	European *Bos taurus*
N'Dama (Guinea)*	36	NGu	West African *Bos taurus*
N'Dama (Guinea Bissau)*	22	NB	West African *Bos taurus*
Somba*	10	SO	West African *Bos taurus*
Gobra*	32	GO	West African hybrids
Sokoto Gudali*	13	SG	West African hybrids
Mbororo*	9	MB	West African hybrids
Kuri*	13	KU	West African hybrids
Anatolian Black*	12	AB	Near Eastern hybrids
Iraqi*	16	IR	Near Eastern hybrids
Afrikaner	14	AF	South African breed
Ankole	14	AK	East African hybrids
Arashie	16	AR	East African hybrids
Ogaden	18	OG	East African hybrids
Watusi	3	WA	East African hybrids
Barotse	16	BA	Southern African hybrids
Kavango	12	KV	Southern African hybrids
Nguni	9	NG	Southern African hybrids
Zebu Malagasy	4	ZM	Madagascar hybrid
Nellore*	12	NE	Indian *Bos indicus*
Ongole*	3	ON	Indian *Bos indicus*

Figure 4.4 *Map showing approximate sampling sites for cattle populations included in the current study. The areas of the pie charts are proportional to the number of animals in the population. The pie charts represent a previously calculated mR value for each breed. This is a measure of Bos taurus/Bos indicus admixture which has been calculated using European Bos taurus and Indian Bos indicus breeds as parental populations. The Nellore breed was sampled in Brazil, but is comprised of Ongole individuals that were brought to South America (*FELIUS *1995).*



4.2.2 SELECTION OF X-CHROMOSOME MICROSATELLITES

15 X-chromosome microsatellites were selected from the Meat Animal Research Center (MARC) database *(http://sol.marc.usda.gov/genome/genome.html)* and previously published literature (DROEGEMUELLER *et al.* 2000). All of the microsatellites selected have been linkage mapped to a 20 cM region on the X-chromosome, except for BMS1820 and ED1MS1, which are 9 cM away from one end of the cluster of microsatellites. Some of

the markers included have also been included in two radiation hybrid maps of the bovine X-chromosome (BTAX). Order of markers is preserved for the most part between all available maps, reconstructions of which are shown in **Figure 4.4** (AMARAL *et al.* 2002; AMARAL *et al.* 2000; BAND *et al.* 2000; SONSTEGARD *et al.* 2001). One marker (BMS2713) was mapped to different locations in two of the radiation hybrid maps. All subsequent analyses that included marker order as a parameter were carried out using both possible positions and it was found that the results were almost identical. For the purposes of this thesis the marker order listed in **Table 4.2** and **Figure 4.6** was used. **Figure 4.6** is a representation of the portion of the X-chromosome that was targeted in this study showing the relative orders of the microsatellite and single nucleotide polymorphism (SNP) markers. The fifteen microsatellite markers were genotyped in the populations marked with an asterisk. A subset of ten microsatellites was genotyped in all populations. Populations marked with an asterisk were genotyped using a radioactive PCR method while the remaining populations were genotyped using an ABI 377™ system. The two methods are described below.

Figure 4.5 *Radiation Hybrid and Linkage maps of the Bovine X-chromosome (BTAX). Radiation hybrid maps I, II and III are taken from Band et al 2000 and Amaral et al 2000, 2002 respectively. Only markers that are involved in the current study were selected from the radiation hybrid maps. The linkage map is from the MARC database. Relevant markers are highlighted by a blue outline.*

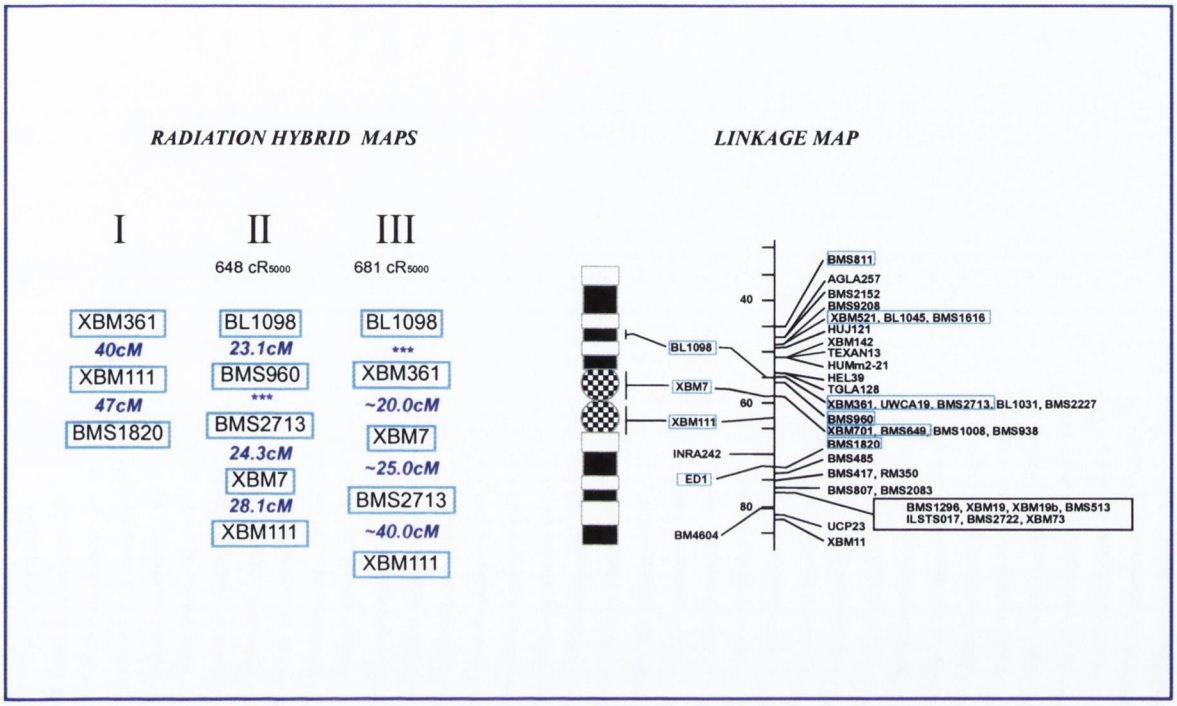
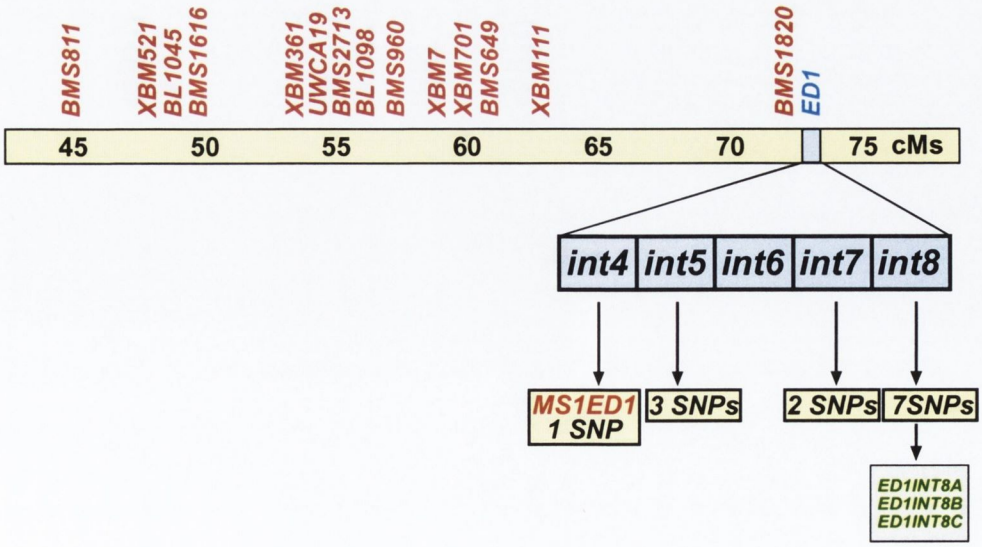


Figure 4.6 *Schematic showing relative order of markers on BTAX. Microsatellite markers are shown in red and the ectodysplasin gene is shown in blue. Novel SNPs that were discovered in introns 4, 5, 7 and 8 are shown in black in the magnified section of the gene. A subset of these, which were genotyped in the current study are shown in green. Centimorgan measures are taken from the linkage map (**Figure.4.4**)*



4.2.3 GENOTYPING OF 15 X-CHROMOSOME MICROSATELLITES

4.2.3.a Radioactive PCR amplification

Primers sequences are listed in **Table 4.2** and were supplied lyophilised by GIBCO BRL and were re-suspended in 1X TE to a stock concentration of 60mM and stored at –20°C. Aliquots were taken from these stock oligonucleotides and diluted to 6.0 μM in 1X TE for use in PCR reactions. Optimal PCR reactions conditions were established using a gradients of annealing temperature and $MgCl_2$ concentration, and are shown in **Table 4.2**. PCR was carried out in 5.0 μl reaction volumes in 96-well polycarbonate microtitre plates. 5 ng of DNA were aliquoted into 96-well polycarbonate microtitre plates and overlaid with 10 μl mineral oil. 4.0 μl of a prepared mastermix was aliquoted into each well. Reaction components are listed in **Table 4.3**.

Table 4.2 *Primer sequences and Optimal PCR conditions for 15 X-chromosome microsatellites. Magnesium Chloride concentration is shown by [$MgCl_2$] and annealing temperature is denoted by T_A*

Locus name	*Accession number*	*Position on BTAX**	*Primers (5'-3')*	*Amplicon size*	*PCR [$MgCl_2$]*	*T_A PCR*
BMS811	G18904	45.0cM	AAAGCTCTCATAAACCAGCACA CAAATTTAGGCCCCTGTAGC	103-121	1.25	59°C
XBM521		49.0cM	CCTTCAGGTCTCAGCTCACC AGTGCTATGAAGAAGACCAGGG	125-131	1.25	59°C
BL1045*		49.0cM	TGCCAGAACAAGTCACAAGC CTCCAAGGGTGTCTCTATCCC	92-100	1.5	59°C
BMS1616*	G18677	49.0cM	CAGTGTGTATAGCATGATTCCG AGTTGGTCTGCATTCATACATTAA	92-116	1.25	54-55°C
XBM361*		50.0cM	ACACACTGAGAATTAAACTACAAAAACC ACACAACTGAAGCCACTTTAAGC	150-162	1.25	58-59°C
UWCA19*	U01794	50.0cM	CTGTGATAACCTATATGGGAAAGGA GAGGACAGCAAAGTGATTCAGT	97-105	1.5	58°C
BMS2713*	G18462	50.0cM	TGAATACCTGTTTCCAGCCC CTCCTAAGTCCAGGAAGCCC	143-155	1.5	56°C
BL1098*		50.5cM	CCACAACTTCCAGAAGCCTC CAGAAACCACCCAAACTAACC	203-235	1.25	56°C

Table 4.2 (cont)

Locus name	*Accession number*	*Position on BTAX**	*Primers (5'-3')*	*Amplicon size*	*PCR [$MgCl_2$]*	*T_A PCR*
BMS960*	G18760	51.5cM	TGTAAATAGCTCCCCTCCTGC CTAGCAATTTGTTGTTCATGGC	106-128	1.75	57°C
XBM7*		58.5cM	CTGTATTAGAGTTCCCTGGAGAAA GCCAACATGCCCTGTAGAAT	180-202	1.5	57°C
XBM701*		58.5cM	TTCCCCTTGAAATCATCTGG TCCAAGTTACCAAAATTGACCC	148-168	2.0	58°C
BMS649*	G18728	58.5cM	CTCGGACTCATAACGCACAT CGAGCAACAAGAGTGAAGGT	116-134	1.75	59°C
XBM111		63.5cM	TGAAAGGGGATGACTTACAACC AATCGCAAAGAGTGGGACAC	129-137	1.25	61°C
BMS1820	G18604	72.5cM	CACAAAGAATCAGACATGATGG ACTGTGTACTGGAGTGGAAAGG	99-103	2.0	58°C
ED1INT8MS1	AJ278907	72.5cM	ACCCAGCCTGTAAATGCTG CCTGCTCTTAAAGAGGTGG	109-139	1.25	45.5°C

Table 4.3 *Table of PCR components for amplification of X-chromosome microsatellites incorporating a radioactive label.*

Reagent	Volume	Final Concentration
ddH20	2.075 µl – 2.2 µl	
10X PLATINUM® *Taq* PCR Buffer	0.5 µl	1X
Forward and Reverse PCR primers	0.375 µl each primer	0.2 µM each primer
dNTPs	0.375 µl	200 µM dATP, dTTP, dGTP; 10.0 µM dCTP
$MgCl_2$	0.125 µl – 0.25 µl	1.25 – 2.5 mM
PLATINUM® *Taq* DNA Polymerase (GIBCO BRL)	0.025 µl	0.25 Units
α-32^P –dCTP (3,000 Ci/µl)	0.025 µl	
Total	4.5 µl	

The reactions were carried out in a MJ Research Peltier Cycler or a Hybaid Omnigene a at 94°C for 4 minutes then 30-35 cycles of 94°C for 45 seconds, 53-59°C for 30-45 seconds, 72°C for 45-60 seconds, then a final extension period of 72°C for 20 minutes. Optimised PCR reaction conditions for the 15 markers are outlined in **Table 4.2**.

4.2.3.b Polyacrylamide Gel Electrophoresis (PAGE) of radioactive PCR product

Upon completion of the PCR reaction 5 µl of formamide-based denaturing buffer containing bromophenol blue and xylene cylanol was added to the completed PCR reaction before it was denatured at 95°C for 3 minutes. The products were transferred immediately to ice and electrophoresed on a SequaGel™ 6% denaturing polyacrylamide gel at 70W for

1.5 – 3 hours. Electrophoresis was carried out using a GIBCO BRL vertical S2 or S2001 gel electrophoresis system (LIFE TECHNOLOGIES). 1X TBE buffer was used for the PAGE. After electrophoresis gels were dried on WHATMAN™ chromatography paper and placed in a cassette with autoradiograph film. After 24-72 hours of exposure, gels were visualised by development of the autoradiograph film. M13 plasmid sequence was run concurrently and was used to size the alleles. All gels were scored manually using sequence and size standards to obtain the correct allele size.

4.2.4 GENOTYPING OF 10 X-CHROMOSOME MICROSATELLITES IN A SUBSET OF SAMPLES

The Southern and Eastern African samples that are not highlighted with an asterisk in **Table 4.1** were genotyped for a subset of ten microsatellites. These markers are marked with an asterisk in **Table 4.2**.

4.2.4.a PCR amplification

Fluorescent labelled versions of the primers listed in **Table 4.2** were obtained from INVITROGEN, with the forward primers end-labelled with the fluorescent dye TET, 6-FAM or HEX. The lyophilised primers were re-suspended in ddH_20, quantified using spectrophotometric methods and diluted to 200ng/µl. Fluorescent dyes used and optimal PCR conditions are outlined in **Table 4.4**. Samples were amplified in 96-well microtitre plates, which included negative controls and samples that had been genotyped using the radioactive PCR method. 10 µl reactions containing 20 ng of DNA, 5 pmol of each primer, 0.25 mM of each dNTP (AMERSHAM), 1 unit of *Taq* polymerase, 1 x PCR buffer (10 mM Tris-HCl pH 8.3) with 1.5 mM MgCl2, 50 mM KCl, 0.001% gelatin (SIGMA), 0.025% Tween-20 (RIEDEL-DE-HAN), 0.025% Nonidet P40 (BHD).

Table 4.4 *Table of optimal PCR conditions and fluorescent dye labels used in the amplification of samples for analysis on an ABI377 automated DNA sequencer.*

Locus Name	*Forward/ Reverse*	*Label*	*PCR [$MgCl_2$]*	*PCR T_A*
BL1045	F	fam	1.50	58
	R	not labelled		
BMS1616	F	tet	1.75	52
	R	not-labelled		
XBM361	F	tet	1.50	55
	R	not labelled		
UWCA19	F	hex	2.00	51
	R	not labelled		
BMS2713	F	hex	1.75	55
	R	not labelled		
BL1098	F	tet	2.00	51
	R	not labelled		
BMS960	F	fam	2.00	52
	R	not labelled		
XBM7	F	hex	1.50	57
	R	not labelled		
XBM701	F	fam	2.50	56
	R	not labelled		
BMS649	F	fam	1.75	57
	R	not labelled		

PCR cycling employed a GeneAmp PCR system 9700 thermal cycler (APPLIED BIOSYSTEMS) with heated lid. Cycling consisted of 95°C for 5 minutes then 40 cycles of

95°C for 45 seconds, 50-59°C for 60 seconds, 72°C for 45-60 seconds, then a final extension period of 72°C for 10 minutes.

4.2.4.b PAGE of PCR products

PCR products were analysed by electrophoresis in 4.25% denaturing polyacrylamide gels on an ABI 377 automated DNA sequencer using the internal size standard GENESCAN 350 TAMRA size standard (APPLIED BIOSYSTEMS). Analysis was carried out using GENESCAN (*v.* 3.1.2, APPLIED BIOSYSTEMS, CA). The third order least squares method was used to size alleles. Results were further analysed using the GENOTYPER software (*v.* 2.0, APPLIED BIOSYSTEMS, CA).

4.2.5 SNP DETECTION AND GENOTYPING

4.2.5.a Amplicon selection and primer design

One of the microsatellites used in the current study is found in an intron of the Ectodysplasin (ED1) gene. This gene has been partially sequenced in a European *B. taurus* individual and represents one of the largest (>37kb) contiguous portions of bovine sequences in GenBank database (Accession Number AJ278907) (DROEGEMUELLER *et al.* 2000). The ectodysplasin gene in *B. taurus* is located on the X-chromosome and is mapped to a region that is very close to the BMS1820 marker (AMARAL *et al.* 2002). The position of the gene is shown in **Figures 4.5 and 4.6**.

There are 5 introns contained in the sequence of over 37kb. The positions of these introns in the GenBank sequence are as follows: 4 (19840-24625bp), 5 (24806-26695bp), 7 (27678-30541bp) and 8 (30673-32563bp). Approximately 1000kb was amplified from each of these introns using oligonucleotide primers designed by eye following published guidelines, and with the assistance of the JELLYFISH software tool

(http://jellyfish.labvelocity.com), (MACHUGH *et al.* 2000). The principal guidelines that were adhered to were:

- The oligonucleotides should be greater than 20 nucleotides long and preferably 22-25bp
- GC content should be 40-60%
- 3' complimentarity and any opportunity for secondary structure to form should be avoided
- A double A residue was incorporated at the 3' end of each primer.
- The sequences and locations of the primers are shown in **Table 4.5**.

Table 4.5 *Table of primer sequences to amplify approximately 1kb from 4 of the introns of the bovine ED1 gene. The primer position shown where the 3' end of the oligonucleotide is found within the GenBank notated intron (AJ278907).*

Amplicon Name	*Amplification primer sequences (5'-3')*	*Primer position*
ED1INT4	AGCTATCCCTCGCCTATTTCTCAA	415
	CCTGCTCATGCTTCTGTGATTGAA	1464
ED1INT5	CAGTGGAGAGCTAGGAATTGAACAA	390
	CCTGTACGACATAAATTCCACGAGAA	1393
ED1INT7	TCAGGAAGCAGCAAGCCAGCCAA	952
	AGGGAACTCAAACCAGGGCTCTGTAA	1807
ED1INT8	CTCCCCACACCACTGGCTTTATAA	410
	TCCCTGCTGCTGAATTCTCAACAA	1343

4.2.5.b PCR amplification of intronic sections

Primers were supplied lyophilised by GIBCO BRL and were re-suspended in 1x TE to a stock concentration of 60mM and stored at –20°C. Aliquots were taken from these stock

oligonucleotides and diluted to 6.0 μM in 1X TE for use in PCR reactions. Optimal PCR reaction conditions were established, using gradients of annealing temperature and $MgCl_2$ concentration, and are shown in **Table 4.6**.

Table 4.6 *Optimal PCR conditions for amplification of 1kb sections of 4 introns from the ED1 gene of B. taurus.*

Amplicon	*PCR [$MgCl_2$]*	*PCR T_A*
Intron 4	1.5	61°C
Intron5	1.25	66°C
Intron7	1.5	70°C
Intron8	2.0	66.5°C

PCR reactions were subsequently carried out in 50 μl volumes using an OmniGene Thermal Cycler (HYBAID) or Peltier Cycler (MJ RESEARCH) following the protocol outlined in **Table 4.7**. Reaction volumes are given for 1 PCR amplification reaction. A mastermix of all reaction components was prepared, excluding the template DNA and scaled to the appropriate number of samples. 48μl aliquots of this mastermix were added to 2μl of the DNA template contained under 10μl mineral oil in a 96-well polycarbonate microtitre plate. Evaporation and cross contamination during aliquoting is prevented by the oil overlay. PCR cycling parameters consisted of an initial denaturation step at 94°C for 3 minutes, followed by 37 cycles of 94°C for 30 seconds, 61°C-70°C for 1 minute and 72°C for one minute, and a final extension at 72°C for 20 minutes. These cycling parameters were modified slightly from those previously described allowing a longer extension time to successfully amplify products approximating 1kb in size. Intron products were successfully amplified from a member of each of the two bovine lineages. An Aberdeen Angus sample (European *B. taurus*) and Indian Ongole (Indian *B. indicus*) were chosen as

these represents the purest sample available based on previous publications and examination of microsatellite data from this region of the X-chromosome (KUMAR *et al.* 2003; MACHUGH *et al.* 1997).

Table 4.7 *Table of PCR components for amplification of 1kb section of introns of ED1 gene of B. taurus.*

Reagent	***Volume***	***Final Concentration***
ddH20	32.75 µl – 32.0 µl	
10X Platinum *Taq* PCR Buffer (GIBCO BRL)	5.0 µl	1X
Forward and Reverse PCR primers	2.5 µl each primer	0.3 µM each primer
dNTPs	3.75 µl	200 µM each nucleotide
$MgCl_2$	1.25 µl – 2.0 µl	1.25 mM
Platinum *Taq* DNA Polymerase (GIBCO BRL)	0.25 µl	2.25 Units
DNA template	2.0 µl	20 – 30 ng
Total	50 µl	

Successful amplifications were confirmed by visualisation on a 6% agarose gel. The products were run with a 100bp size ladder and contained 0.25µM ethidium bromide, which interchelates with DNA and is illuminated by *UV* radiation. PCR products were purified using a Concert™ PCR purification kit (LIFE TECHNOLOGIES). MWG BIOTECH supplied commercial automated double-stranded DNA sequencing of amplicons.

4.2.5.c SNP detection and Single Strand Conformational Polymorphism (SSCP) primer design and PCR amplification

Single Nucleotide Polymorphisms (SNPs) and indels were identified by aligning the sequences obtained from each intron for the two individuals and the database sequences using the ClustalW program (HIGGINS and SHARP 1988; THOMPSON *et al.* 1994). Mutations were confirmed by examination of the DNA sequence traces provided by MWG BIOTECH. The amplicons from introns 4, 5, 7 and 8 contained 1, 3 2 and 7 SNPs respectively.

Intron 8 of the ectodysplasin gene contained the most SNPs therefore primers were designed to amplify sections of around 300-500bp which could be used for large scale genotyping using the single strand conformational polymorphism method (SSCP) (ORTI *et al.* 1997). Internal primers were designed and obtained as described above and are listed in **Table 4.8**. These primers were combined with the original primers designed to amplify the longer portion of intron 8 to produce three shorter amplicons, each containing at least one SNP. These were denoted ED1INT8A, ED1INT8B and ED1INT8C.

Table 4.8 *Table of primer sequences to amplify approximately 300-500 bp of intron 8 of the bovine ED1 gene. The primer position shown where the 3' end of the oligonucleotide is found within the GenBank notated intron 8 (AJ278907).*

Amplicon Name	*Amplification primer sequences (5'-3')*	*Primer position*
ED1INT8A	CTCCCCACACCACTGGCTTTATAA	410
	TGGGCTCAGCAAATGCTCAA	782
ED1INT8B	TCCCAGGAGAGTAGTCAGAGAA	751
	CTTGGTGAATAGAAGAAGTGCTAA	1206
ED1INT8C	CAGTGTCCTATGGGACATTTGTCCTAA	1001
	TCCCTGCTGCTGAATTCTCAACAA	1343

PCR-SSCP analysis was carried out using a modified version of previously published protocols (LAGZIEL *et al.* 1996; SUNNUCKS *et al.* 2000; TERES *et al.* 1996). The populations marked with an asterisk in **Table 4.1** were used for this analysis. Radioactively labelled nucleotides were incorporated into the amplicons during the PCR reaction to enable visualisation after subsequent PAGE. PCR reactions were carried out in 10.0µl volumes in 96-well polycarbonate microtitre plates, using an OmniGene Thermal Cycler (HYBAID) or Peltier Cycler (MJ RESEARCH). Each well in the plate contained 1.0 µl of each DNA sample covered by a 10.0 µl mineral oil overlay. 4.5 µl of mastermix was aliquoted into every well, including negative controls. The previously sequenced individuals were included to aid visualisation of results. The PCR reaction components are described in **Table 4.9**. Optimised PCR and electrophoresis parameters for each PCR-SSCP amplicon are given in **Table 4.10**. Thermocycling was conducted as follows: 94°C for 3 minutes, followed by 35 cycles of 94°C for 30 seconds, 60°C-65°C for 40 seconds and 72°C for one minute, and a final extension at 72°C for 20 minutes.

Table 4.9 *Table of PCR components for amplification of 300-500 bp sections of intron 8 of ED1 gene of B. taurus.*

Reagent	*Volume*	*Final Concentration*
ddH20	5.675 µl – 5.925 µl	
10X Platinum *Taq* PCR Buffer (GIBCO BRL)	1.0 µl	1X
Forward and Reverse PCR primers	1.0 µl each primer	0.3 µM each primer
dNTPs	0.75 µl	200 µM dATP, dTTP, dGTP; 10.0 µM dCTP
$MgCl_2$	0.25 µl – 0.5 µl	1.25 mM
Platinum *Taq* DNA Polymerase (GIBCO BRL)	0.05 µl	0.5 Units
α-32^P –dCTP (3,000 Ci/µl)	0.025 µl	
DNA template	1.0 µl	20 – 30 ng
Total	10 µl	

Table 4.10 *Optimal PCR conditions for amplification of 300-500 bp sections of intron 8 of ED1 gene of B. taurus.*

Amplicon	*PCR [$MgCl_2$]*	*PCR T_A*	*Electrophoresis Conditions*
ED1INT8A	2.5	62°C	7W 15hrs 4°C
ED1INT8B	1.75	60°C	7W 20hrs 4°C
ED1INT8C	2.0	65°C	7W 13.5hrs 20°C

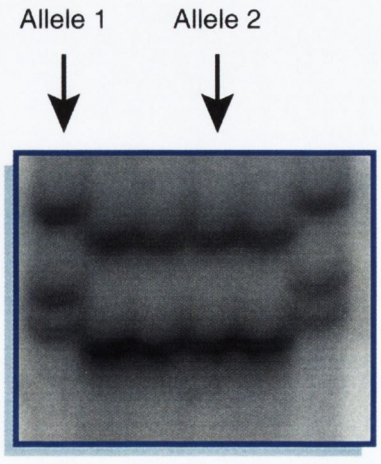
4.2.5.d Gel Electrophoresis of PCR-SSCP products

10.0 µl of formamide-based loading buffer was added to each radioactive PCR product. These were then heat-denatured at 95°C for 3 minutes and quickly placed on ice. 2.5µl -of each samples was loaded on a 5% non-denaturing polyacrylamide gel (AccuGel™,

NATIONAL DIAGNOSTICS) and electrophoresis was carried out using a GIBCO BRL vertical S2 or S2001 gel electrophoresis system (LIFE TECHNOLOGIES). 1X TBE buffer was used for the PAGE. Electrophoresis was carried out at both 20°C and 4°C as temperature can affect the conformation of DNA in the gel. Certain SNPs only form differing conformations that are detectable by the SSCP technique at one or other temperature. After electrophoresis gels were dried on WHATMAN™ chromatography paper and placed in a cassette with autoradiograph film or an imaging plate. After 48-72 hours of exposure, gels were visualised either by development of the autoradiograph film, or reading the imaging plate using a FUJIFILM FLA-3000 Phosphoimager.

Band shifts were scored by eye using previously sequenced individuals as controls. An example of a SSCP band shift is shown in **Figure 4.7**. Any individual that had a band shift that was not identical to one of the previously sequenced controls was sent for automated sequencing and the putative SNP was refuted or validated by examination of the DNA sequence traces provided by MWG BIOTECH.

Figure 4.7 *SSCP band shift on polyacryamide gel visualised by phosphoimager. This band shift represents an SNP that is found in the ED1INT8C amplicon, which is located in intron 8 of the bovine ectodysplain gene.*



4.2.6 ANALYSIS OF MICROSATELLITE DATA

4.2.6.a Allele frequencies

Allele frequencies were calculated by the MICROSATELLITE TOOLKIT (PARK 2001) and are in **Appendix A**.

4.2.6.b Phylogeography

The allele sharing model of microsatellite mutation was used in place of stepwise mutation model as the highly polymorphic nature of the microsatellites under study introduced large amounts of noise into phylogeographic analysis. Population pairwise F_{ST} values (**Equation 3.1**) were calculated as an estimate of genetic distance between breeds using ARLEQUIN *v.* 2.0 (SCHNEIDER *et al.* 2000). Multidimensional scaling (MDS) analysis of population pairwise F_{ST} values between breeds was carried out in SPSS (*v.* 11.0 SPSS Inc. Chicago, IL). As explained previously, (**Chapter 3**), MDS does not impose ancestral status upon the populations under consideration and instead summarises affinities among populations based on a distance measure that has been calculated between them. In addition, each axis may exhibit independent genetic trends. An inter-haplotypic distance matrix of individual chromosomes was constructed in ARLEQUIN. The distance matrix was constructed using the allele-sharing method to count numbers of differences between haplotypes. Trees of individual haplotypes were constructed using NEIGHBOR from the PHYLIP analysis package (FELSENSTEIN 2001).

4.2.6.c Mismatch distributions

Mismatch distributions for each regional population were calculated in ARLEQUIN using both the alleles sharing and the stepwise mutation models (SMM) to count the number of differences between individuals within each regional grouping. Permutation testing was

carried out to examine the validity of assigning each individual to a regional population. Individuals were randomly assigned to a population and this population was compared to the assigned populations. This was repeated 1000 times. Of all these permutations the assigned geographical divisions belong to a group of permutations of <0.1%, which explains a maximum amount of variance.

4.2.6.d Linkage Disequilibrium

Pairwise linkage disequilibrium between locus pairs was carried out using an extension of Fisher's exact test in ARLEQUIN and the GENEPOP program (SLATKIN 1994, SLATKIN & EXCOFFIER 1996). A contingency table was built whereby the η_1 x η_2 entries of the table are the observed haplotype frequencies. η_1 and η_2 are the number of alleles at locus 1 and locus 2 respectively. The test calculates the probability of finding a table with the same marginal totals and which has a probability less than or equal to the observed table. The null hypothesis is that there is no association between the two loci. A Markov chain, or random walk in probability space, is used to efficiently explore the probability space of all possible tables with the same marginals and avoids the need to calculate all possible contingency tables. This was done with 10000 dememomorisation steps to ensure that the chain starts from an independent starting point. The *p*-value of the test is the proportion of visited tables having a probability smaller or equal to that of the observed contingency table (SCHNEIDER et al. 2000). No attempt was made to correct for multiple testing, as the large number of tests would result in low statistical power.

The ARLEQUIN system requires a cut-off *p*-value to be pre-specified and locus pairs that are associated more that the pre-assigned cut off point are shown in results. Two levels of significance were used: p=0.05 and p=0.001. This was used for the analysis shown in **Figure 4.10** to demonstrate the difference between significant and highly significant

associations over the 15 loci on the X-chromosome. The GENEPOP software outputs *p*-values for all pairwise locus combinations and was used in the comparison between the physically linked X-chromosome markers and the non-linked autosomal markers in **Figure 4.11 and 4.19**. Unlinked microsatellite data was taken from previously published studies (HANOTTE *et al.* 2002; MACHUGH *et al.* 1997; MEGHEN *et al.* 1994). The number of associations with *p*-values less than 0.05 and 0.01 were counted as significant and highly significant respectively. The number of pairwise associations that were significant at these two levels was counted for both linked and non-linked markers.

4.2.6.e Allele ancestry

Both a fully Bayesian and a partially Bayesian approach were used to assign ancestry to the X-chromosome microsatellite alleles. The fully Bayesian approach was implemented using the linkage model of the program STRUCTURE (*v.* 2.0) (FALUSH *et al.* 2003; PRITCHARD *et al.* 2000a). Map distances between loci are input with allelic data and probable ancestry is assigned to each allele in each haplotype. This model takes both individual ancestry and allele frequencies in putative parental populations into account when assigning allele ancestry. The program was run with a burn-in period of 100,000 repetitions of the Markov chain and then 1,000,000 repetitions for data collection. As explained in **Chapters 2 and 3**, the program assumes original parental populations in Hardy Weinberg equilibrium and assesses LD that is produced during admixture.

Two models were tested using this method of ancestry assignment, one assuming 2 ancestral populations (k=2 model) contributing to the present day admixed populations, and one that assumes that there were three original contributors (k=3 model). The k=2 model clearly separates the *B. taurus* and the *B. indicus* lineages, while the k=3 model further separates the taurine lineage into European/Near Eastern types and African types.

The algorithm estimates a probability of an allele being derived from each of the inferred parental populations. The model assumes that every allele in the population is derived from one of the inferred parentals, therefore the probability values for each individual allele sum to unity. Each chromosome was reconstructed diagrammatically using allele ancestry probability to assign genealogical origin to the section of the X-chromosome under study. The mapped order of the microsatellite markers was used and probability of ancestry was examined at values of 0.8 and a less stringent 0.5. Each individual haplotype was inspected to see how intact parental haplotypes might be in either of the two hybrid populations.

The partially Bayesian approach ignores individual admixture, but bases calculations on parental allele frequencies and breed admixture coefficients as priors. Loci are treated independently and *mR*, from the ADMIX program (BERTORELLE and EXCOFFIER 1998), is used as the admixture coefficient. In this case pre-assigned, as opposed to inferred, parental populations were used; this statistic bears some similarity to Chakraborty's gene identity coefficient (CHAKRABORTY 1975; CHAKRABORTY *et al.* 1992). An Indian *B. indicus* sample was used to represent the putative indicine ancestor while a pooled sample of the European and West African *B. taurus* individuals was used to represent a putative taurine ancestor. These may be designated ancestral populations 1 and 2 respectively. The probability of an allele having *B. indicus* ancestry was calculated by implementing Bayes theorem as follows:

$$p(Z|A) = \left(\frac{p(A|Z)p(Z)}{p(A|Z)p(Z) + p(A|T)p(T)} \right) \qquad \textbf{Equation 4.1}$$

where $p(Z|A)$ is the probability of an allele having ancestry from ancestral population 1 given the data, $p(A|Z)$ is the probability of the allele in question being from parental

population 1, $p(Z)$ is the probability of any allele in this location coming from parental population 1, $p(A|T)$ is the probability of the allele in question coming from parental population 2 and $p(T)$ if the probability of any allele in this location coming from parental population 2. $p(A|Z)$ is the breed population 1 admixture coefficient (mR, see **Figure 4.4**) from previously published literature, and $p(Z)$ is the frequency of the allele in parental population 1. $p(A|T)$ will be 1-mR and $p(T)$ will be the frequency of the allele in parental population 2.

As with the fully Bayesian approach, individual chromosomes were visualised diagrammatically by probability of ancestry from either the *B. taurus* or the *B. indicus* lineage probabilities of 0.8 and 0.5. In this case, however, only 2 parental populations were modelled, equivalent to the fully Bayesian k=2 model.

The number of possible recombinations between the *B. taurus* and *B. indicus* lineages that may be detected within each individual chromosome was estimated by direct counting of adjacent alleles with different assignments. The 0.5 probability of allele ancestry was used and each chromosome was scored based on the number of recombinations. In the case of the fully Bayesian k=3 model, alleles that were assigned to either of the 2 inferred parental populations characterised by European taurine and West African taurine were considered to be taurine alleles. Alleles that were not assigned were not included. These recombination scores were averaged over breeds. The analysis was carried out for the 2 fully Bayesian models (k=2 and k=3) and the partially Bayesian model.

The number of recombinations estimated by the partially Bayesian model was plotted against the admixture coefficient of each breed (mR, or 1-mR if mR>0.5) and a regression

analysis was carried out in SPSS (*v.* 11.0). Regression residuals were plotted with sampling location using the ARCVIEW GIS software package (Environmental Systems Research Institute, ESRI, Redlands, CA).

4.2.7 ANALYSIS OF SNP DATA

4.2.7.a Allele frequencies

Allele frequencies for the 3 SNP markers were calculated using the MICROSATELLITE TOOLKIT (PARK 2001) and are shown in **Figure 4.25**.

4.2.7.b Allele ancestry

SNP data were combined with the data from the microsatellite ED1MS1 as these markers are all known to be located on one contiguous stretch of DNA, within a region smaller than 37kb. This analysis was carried out for the breeds marked with an asterisk in **Table 4.1**. The fully and partially Bayesian approaches to assigning alleles to parental populations to detect recombinations was repeated for these four markers.

The fully Bayesian approach was implemented using the linkage model of the program STRUCTURE (*v.* 2.0) with the kilobase distances between loci being included in the input. The program was run with a burn-in period of 100,000 repetitions of the Markov chain and then 500,000 repetitions for data collection. 2 parental populations were inferred, therefore 2 probability values were estimated for each allele. These probabilities sum to 1 and each estimates the likelihood of the allele being from one of the 2 inferred parental populations. The partially Bayesian approach employed **Equation 4.2** to assign ancestry probabilities to each of the 4 alleles. Previously calculated *mR* values (**Figure 4.4**) were used as breed admixture coefficients.

The region of the X-chromosome containing the 4 markers was reconstructed graphically using likely ancestry of alleles. The number of crossovers was estimated by direct counting. Any alleles that could not be assigned to an ancestral population with a

probability of greater than 50% were excluded from the analysis. The number of recombinations estimated by the partially Bayesian model was plotted against the admixture coefficient of each breed (*mR*) and a regression analysis was carried out in SPSS (*v.* 11.0).

4.3 RESULTS

For clarity, the results section is divided into three sections. The first section deals with analysis carried out on the thirteen breeds that are indicated by as asterisk in **Table 4.1**. These populations were genotyped for 15 microsatellite markers and three single nucleotide polymorphisms (SNPs). The second results section covers the analysis conducted including all breeds listed in **Table 4.1**. These were each genotyped for a subset of ten of the microsatellite markers. Linkage disequilibrium analysis in the second results section will include only those breeds NOT marked with an asterisk in **Table 4.1**. For some of the following analyses the breeds are bracketed together according to the regional group to which they are assigned in **Table 4.1**. The final section deals includes analysis of the newly discovered SNP markers.

4.3.1 RESULTS 1

Dataset 1 – 15 markers in 13 breeds

4.3.1.a X- chromosome maps

15 of the microsatellite markers were successfully amplified in the five geographic metapopulations (n=240). The map position of the markers was designated using the MARC linkage map (see Materials and Methods) and a combination of two RH_{5000} maps (AMARAL *et al.* 2002; AMARAL *et al.* 2000; BAND *et al.* 2000). A composite of these two maps and the linkage map are shown in **Figure 4.5**. There is good concurrence between the two maps with regard to marker order. All of the following analyses were carried out using the markers assuming the order shown in **Figure 4.6**. It can be seen that most of the markers are localised in an area spanning approximately 20 cM (BMS1820 and ED1MS1 are outside this area), but that many of them may be more tightly clustered and are

currently mapped to the same location. This should facilitate detection of both longer and shorter-range linkage effects.

4.3.1.b Allele frequencies

The frequencies of the 15 microsatellites that were typed have been outlined in **Appendices A5 and A6**.

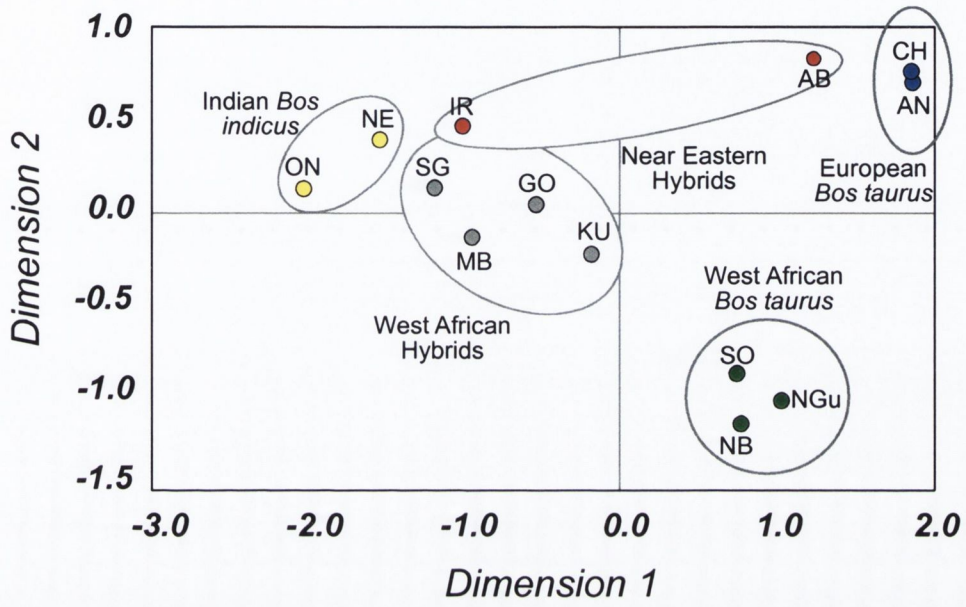
4.3.1.c Phylogeography

As these markers are all located on a section of the X-chromosome, it is possible that linkage effects could distort any phylogenetic information that they contain. Therefore, phylogeography at the breed and individual level is examined.

4.3.1.c.i Population Level

Inter-population F_{ST} genetic distances were also represented in 2-D following multi-dimensional scaling. The MDS plot of population pairwise F_{ST} values demonstrates the resolving power of the markers in the study and displays the major phylogeographic trends in the data (**Figure 4.8**). Breeds from proximal geographic locations with similar genetic histories are found to cluster. The first dimension of the plot clearly separates the two main cattle lineages, while the second divides European/Near Eastern *B. taurus* from African taurine. This supports previous principal component analysis (MACHUGH *et al.* 1997). The two Near Eastern breeds that are included have quite different admixture backgrounds (LOFTUS *et al.* 1999), explaining why the Anatolian Black groups strongly with the European breeds, while the Iraqi is closely related to the African *B. taurus*/*B. indicus* hybrids. However, their intermediate position reflects partial indicine ancestry as indicated by admixture analysis.

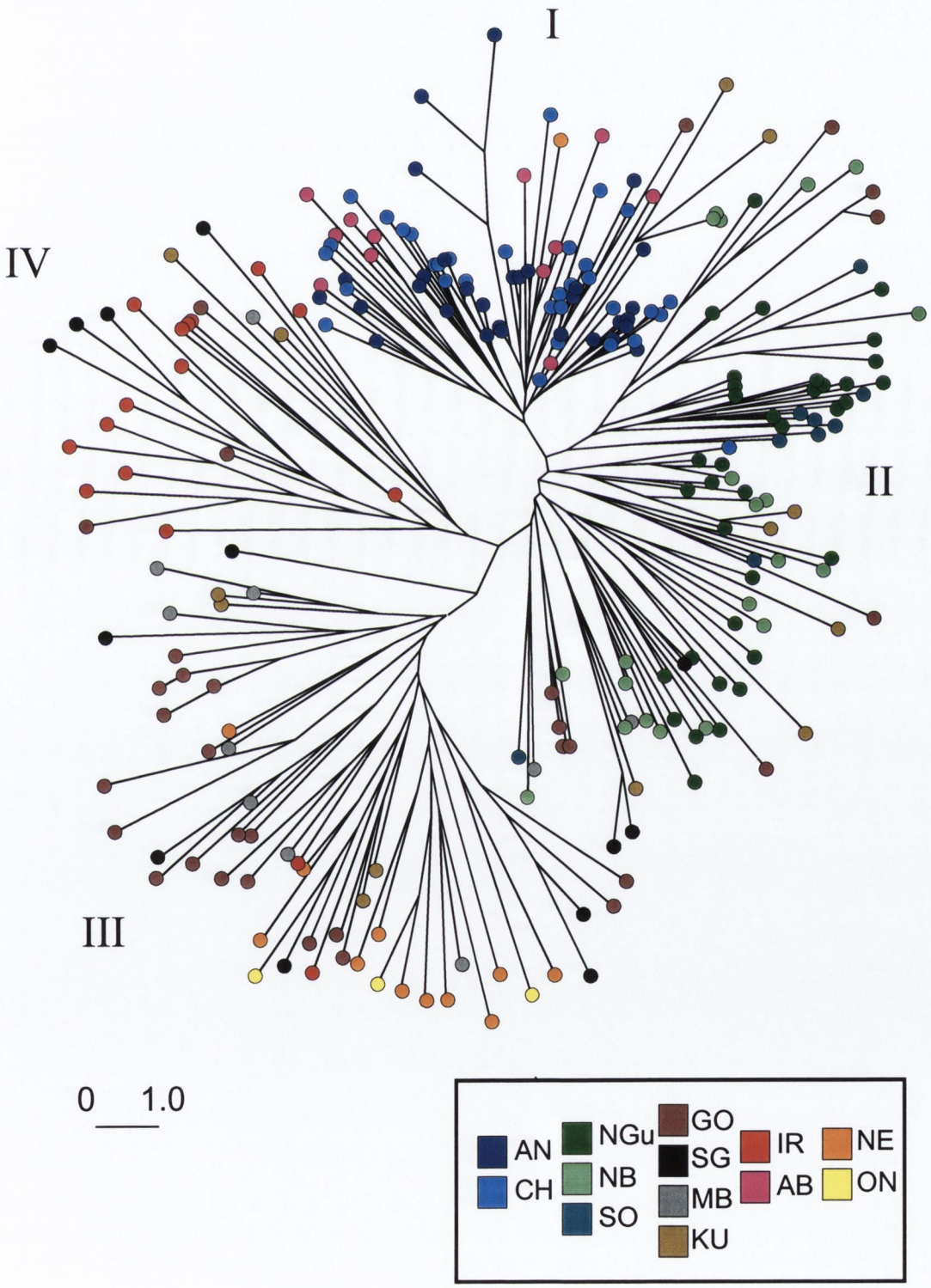
Figure 4.8 *Multi-dimensional scaling plot of population pairwise F_{ST} values calculated from 15 X-chromosome microsatellites. The first dimension is divided between B. taurus and B. indicus. The second dimension (which represents a smaller distance) separates the two taurine groups. The grey lines demonstrate that breeds from the same geographical area and genetic background cluster together. Breed codes are given in* ***Table 4.1.***



4.3.1.c.ii Individual level

Each chromosome haplotype was treated as an operational taxonomic unit (OTU), and an allele sharing distance matrix was used to construct an unrooted neighbour-joining tree. The resulting tree shows that most chromosomes from a particular geographic region with a similar hybrid background cluster together (**Figure 4.9**). There are, however, some haplotypes that appear to fall within other branches of the tree. They demonstrate that while the average number of distances between individuals within breeds (as illustrated in **Figure 4.8**) may be as expected from previous phylogeographic knowledge, the individual level may be different when only a small chromosomal region is considered. Regional groups are highlighted by roman numerals, and it can be seen that most branches of the tree contain chromosomes from exclusively one regional grouping.

Figure 4.9 *Unrooted neighbour-joining tree of individual X-chromosome haplotypes constructed from alleles sharing differences between 15 microsatellites for 240 individuals. Each circle represents one individual and the colour of the circle shows to which breed the individual belongs. The tree can be separated into 4 main groups, which are denoted by roman numerals. Clade I contains a majority of European Bos taurus individuals, group II, a series of closely related clades, is predominantly composed of non-hybrid African individuals, clade III contains most of the Indian samples and some African hybrids and clade IV consists of African hybrids and members of the Iraqi breed.*

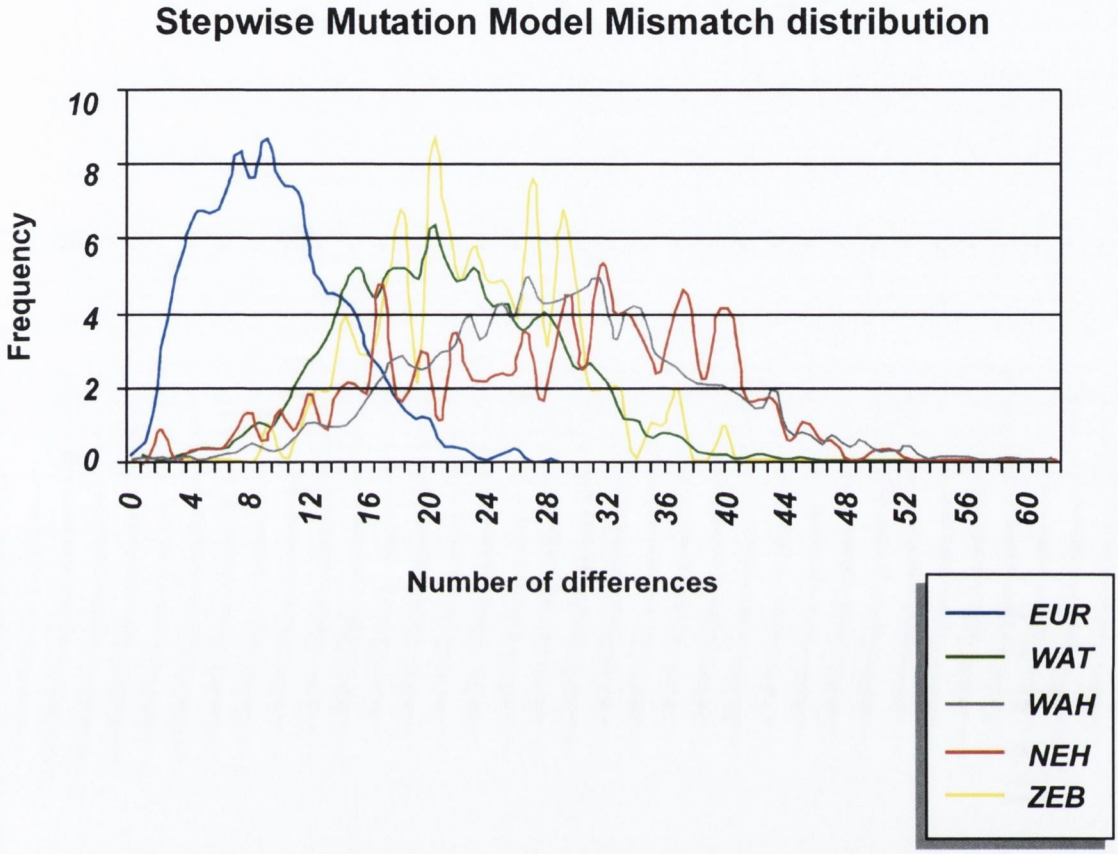


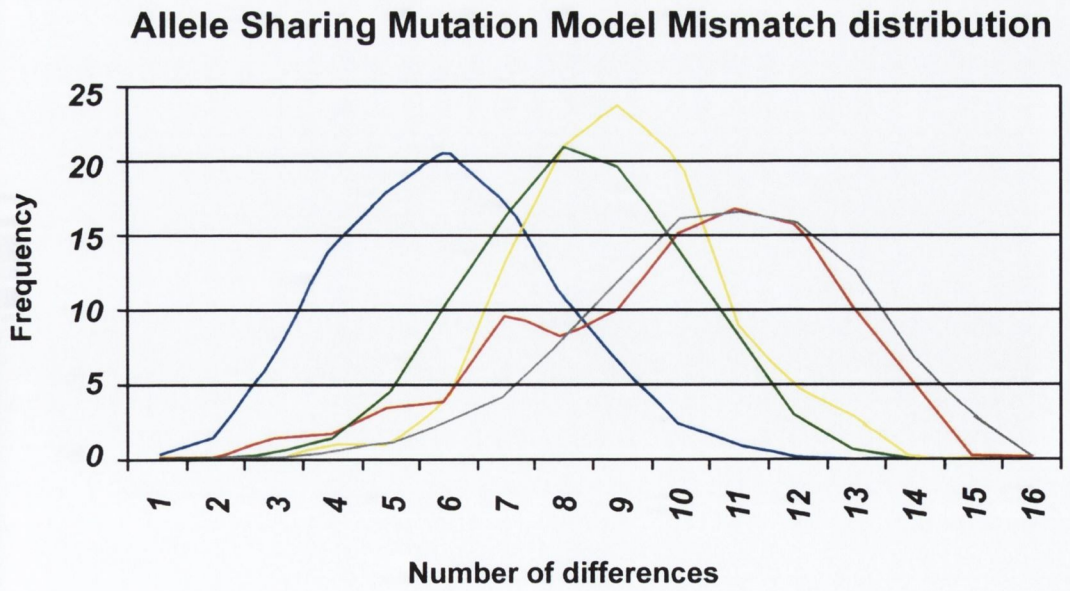
4.3.1.d Mismatch Distributions

A mismatch distribution is a plot of the frequencies of the observed numbers of differences between pairs of haplotypes in a population. While typically utilised with DNA sequence data, here mismatch distributions plot the number of pairwise differences between microsatellite haplotypes within the five regional groupings specified in **Table 4.1**. The distributions were plotted using both the number of differences calculated based on the allele sharing and stepwise mutation models of microsatellite mutation and are shown in **Figure 4.10**. There is such a range of differences between individuals within groups when the stepwise mutation model (SMM) is used (upper graph) that a very ragged distribution is produced. However, when this is compared to the mismatch distribution in the lower graph, which was constructed using allele sharing methods, it is evident that the same general pattern appears in both charts. The curves are all unimodal, which suggests that any hybridisation event may not have occurred in the very recent past. Moreover, this further confirms our regional population groupings.

In the case of both microsatellite mutation models, the modes of the mismatch distributions follow the same pattern. The European *B. taurus* populations possess the lowest number of within-populations differences, typical of an established, older population, while the two hybrid populations have the highest number. Given the hypothesis outlined in **Figure 4.3**, it seems reasonable that these hybrid populations have a greater number of within-population differences because their haplotypes are comprised of divergent genetic material from two source populations. On the other hand, Europe was populated with cattle that had been initially domesticated in the Near East and hence represent a subset of the variation that was originally present in that centre of domestication (TROY *et al.* 2001). It is noteworthy that the distributions are almost identical for the Near Eastern and West African hybrid groups.

Figure 4.10 *Mismatch distributions calculated using the two models of microsatellite evolution. The Y-axis represents the number of individuals separated by the number of differences specified by the X-axis. The legend abbreviations are as follows: EUR=European B. taurus, WAT=West African B. taurus, WAH=West African hybrids, NEH=Near Eastern Hybrids and ZEB=Indian B. indicus.*

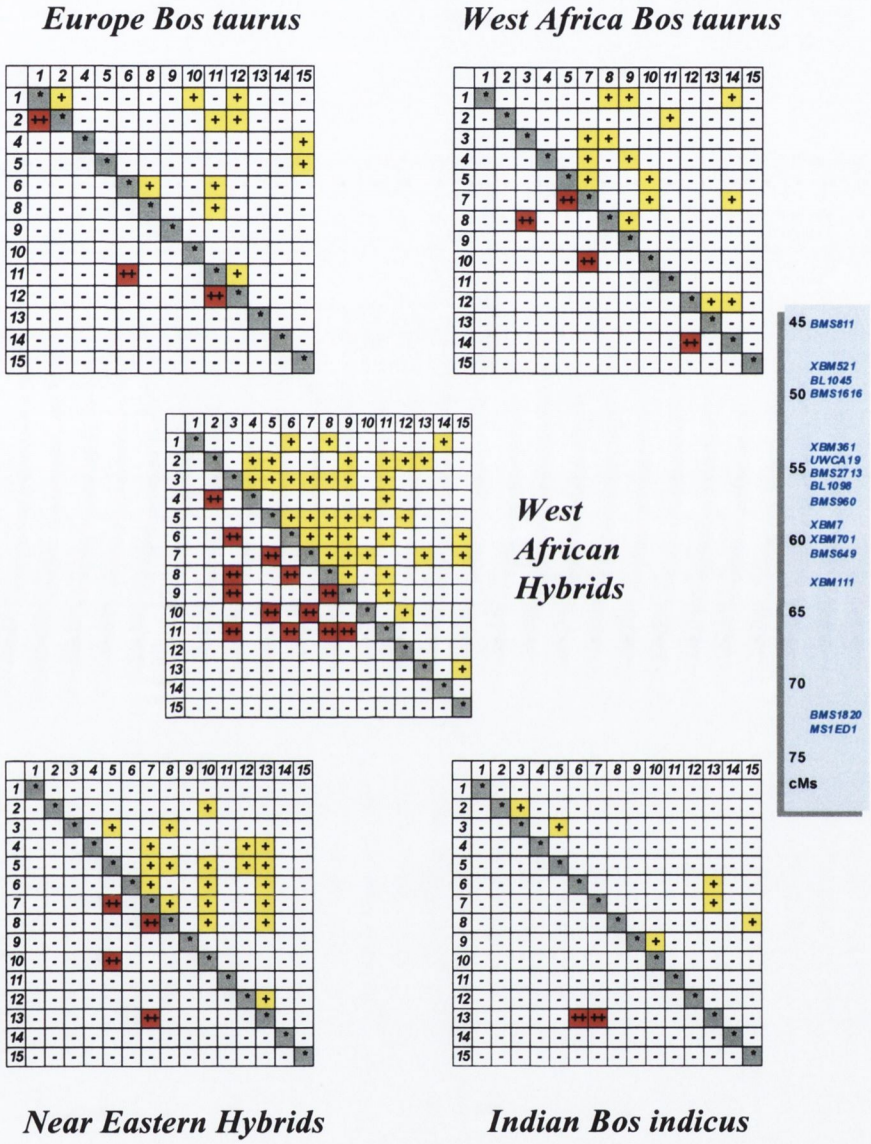




4.3.1.e Linkage disequilibrium between linked markers

Linkage disequilibrium (LD) between the fifteen linked microsatellites on the X-chromosome was assessed using two levels of significance in the five regional populations (p=0.05 and p=0.001). The results are shown in **Figure 4.11**. Each triangular matrix shows all possible pairwise comparisons of the 15 loci in the order that they are found on the chromosome according to the map data available. The top triangle of the matrix shows significant LD when test significance is p=0.05. The lower section of the matrix represents the test done with a significance level of p=0.001. As most of the markers are tightly linked and found within 20 cM of each other fairly high levels of LD would be expected. Homoplasy may also cause some spurious associations to be detected. All of the populations exhibit some disequilibrium in this chromosomal area at a significance level of 0.05. However when the stringency of the test is increased to 0.001 most of the LD ceases to be significant, except within the West African hybrid population where there is still widespread and substantial disequilibrium ranging over eight markers (~9 cM).

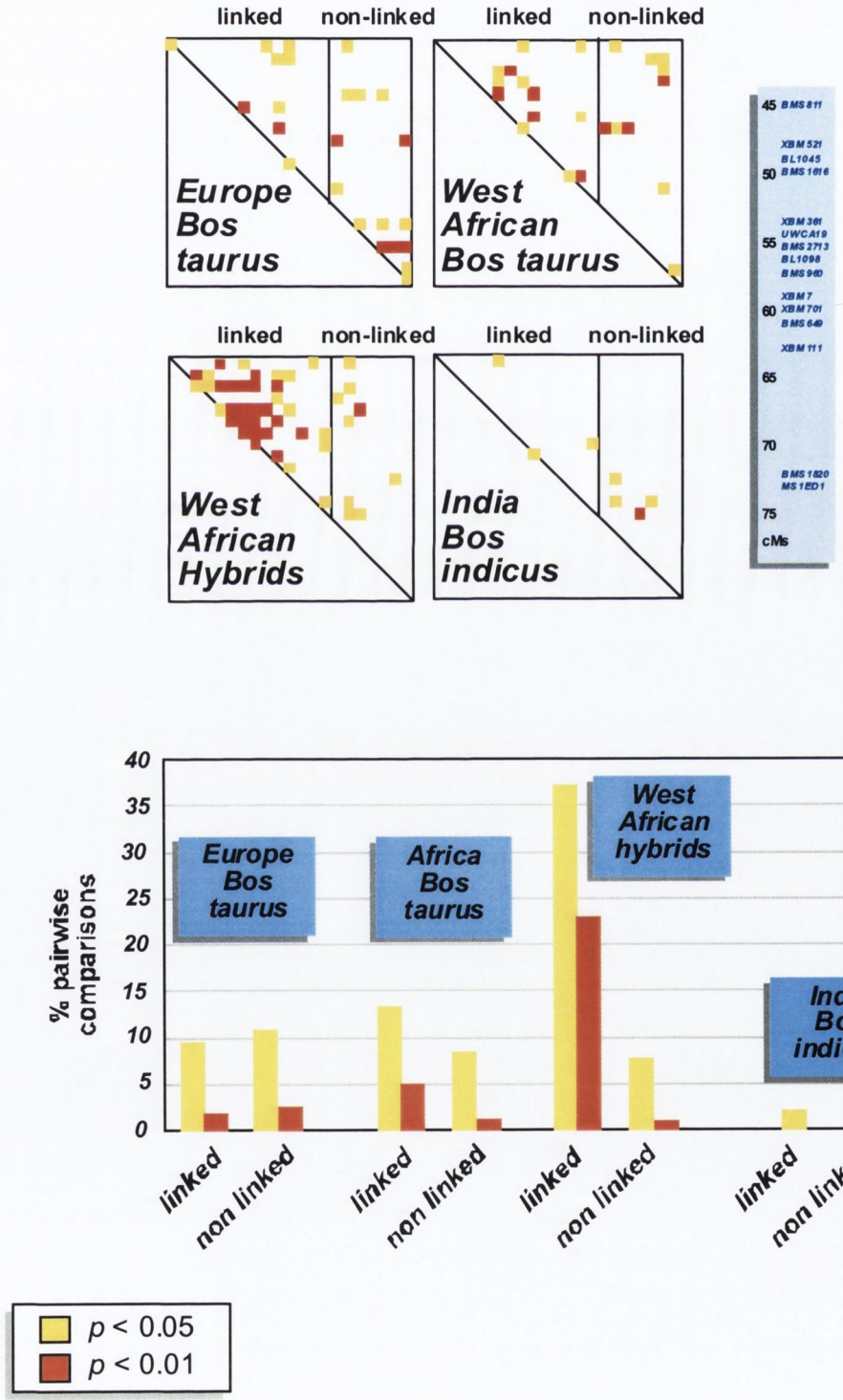
Figure 4.11 *Pairwise linkage disequilibrium between linked X-chromosome microsatellite markers. The matrix shows all possible pairwise comparisons of polymorphic loci in the five regional groupings classified in* ***Table 4.1****. The markers are shown in the order they are mapped on the chromosome. A coloured block indicates a significant pairwise comparison. Results are shown for significance levels p=0.05 (upper triangle, yellow) and p=0.001 (lower triangle, orange). A scale of genetic distances between the microsatellite markers is shown on the right.*



4.3.1.f Linkage disequilibrium between linked and non-linked markers

Additional data from seven autosomal microsatellites was available for all of the regional groupings except for the Near Eastern hybrids. This data was combined with the data from the linked X-chromosome microsatellites and the differences in LD between linked and non-linked markers were examined. The results of this analysis are shown in **Figure 4.12**.

Figure 4.12 *Pairwise linkage disequilibrium results for linked and non-linked markers. The matrix plots are reduced versions of the type of plot shown in* ***Figure 4.11****. They illustrate pairwise associations between markers at significance levels of p<0.05 (yellow) and p<0.01 (red). The markers run from left to right along the top of the plot (numbers 2-22) and from top to bottom along the left-hand side of the plot (numbers 1-21). The linked X-chromosome markers are ordered according to their mapped locations which are shown in the right hand scale . The histograms show the number of differences at the different significance levels for the linked and non-linked markers.*



4.3.1.g Fully and Partially Bayesian Analysis of Allele Ancestry.

Allele ancestry was estimated using two methods: a fully Bayesian analysis was implemented using the program STRUCTURE, and a partially Bayesian analysis was implemented using **Equation 4.1**. The fundamental difference between the two methods is the priors that are employed. The partially Bayesian approach treats the loci independently, and uses a previously calculated admixture coefficient, while the fully Bayesian method simultaneously estimates an admixture coefficient, and uses neighbouring allele ancestry as one of the priors (D. FALUSH, personal communication). In the first approach both two and three ancestral populations were inferred, while the second approach only models the main division in domestic cattle, the indicine and taurine division. Both approaches allow a probable ancestry assignment for each allele in each X-chromosome haplotype. Each X-chromosome haplotype is reconstructed diagrammatically using these assignments.

The results of the Bayesian approach are shown in **Figure 4.13**. The k=2 model plot shows that all alleles in the two European taurine breeds and the West African taurine breeds are strongly derived from a pure *B. taurus* ancestor. Most of the alleles in the Indian *B. indicus* breeds also have a strong probability of having come from an indicine ancestor. Taurine alleles are present in some of the Nellore individuals, but this is not unexpected as there is evidence for limited introgression in this breed, which was sampled in Brazil (KUMAR *et al.* 2003). The hybrid individuals possess alleles from both lineages, but it is difficult to distinguish breeds which have more recombinations than would be expected given their admixture coefficients.

A third element is introduced into the analysis in the k=3 model. Here it is clear that there is separation of the Near Eastern/European and the African *B. taurus* lineages, as was seen

with the second dimension of the MDS plot in **Figure 4.8**. The vast majority of the alleles in the two taurine groups are designated as having come from two different source populations. Unsurprisingly, it can be seen that most of the taurine alleles that are contained within the African hybrid chromosomes are of African origin, while the taurine alleles that the Near Eastern breeds include are predominately of European/Near Eastern ancestry.

The X-chromosome haplotypes that were reconstructed schematically from ancestry probabilities of the partially Bayesian analysis are shown in **Figure 4.14**. This model produces a similar picture to the fully Bayesian k=2 model. However, because a breed ancestry coeffiecient that has been calculated from randomly dispersed microsatellites is used as opposed to the individual ancestry coefficient that is estimated from this chromosomal region, all alleles in the designated parentals are uniformly pre-assigned. The fully Bayesian approach has no designated parental populations and infers these, whereas the partially Bayesian method uses pre-assigned representatives for the ancestral populations. The partially Bayesian approach does not simultaneously estimate either breed or individual admixture levels, whereas the STRUCTURE-based analysis produces posterior distributions for both of these parameters.

The second major difference between the approaches is that the fully Bayesian method then uses individual admixture level as a prior for estimating allele ancestry, so increasing the likelihood that adjacent alleles will be designated as having probably derived from one ancestral population. The same allele may be designated as having come from different parental populations in different individuals. The fully Bayesian approach also takes into account the probable ancestry of the surrounding alleles. This approach decreases any

chance of detecting recombination and certainly represents a very conservative way of identifying putative crossovers. The partially Bayesian approach uses breed admixture coefficient as a prior (as estimated by *mR*), but ignores both individual and adjacent allele ancestry.

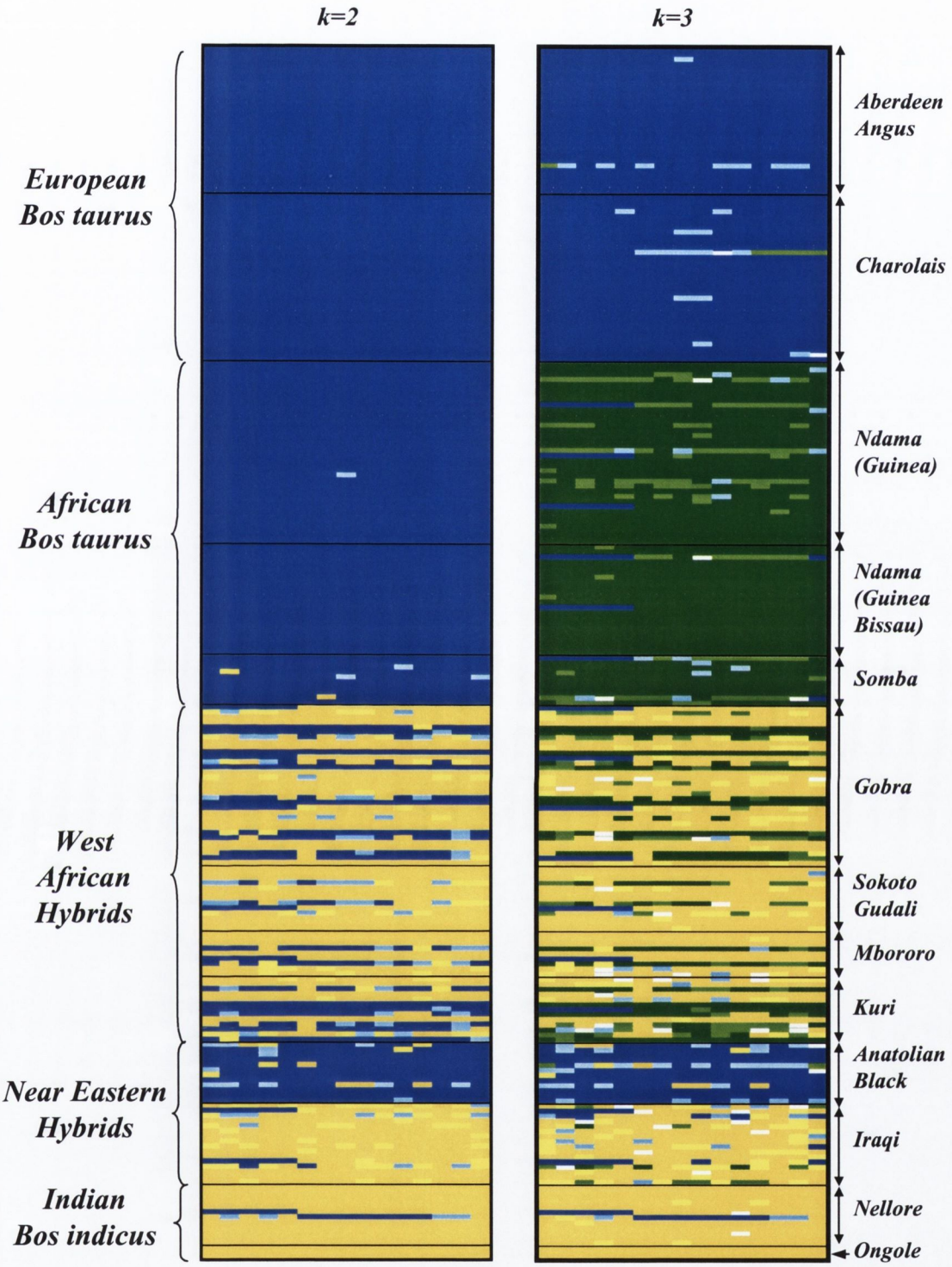
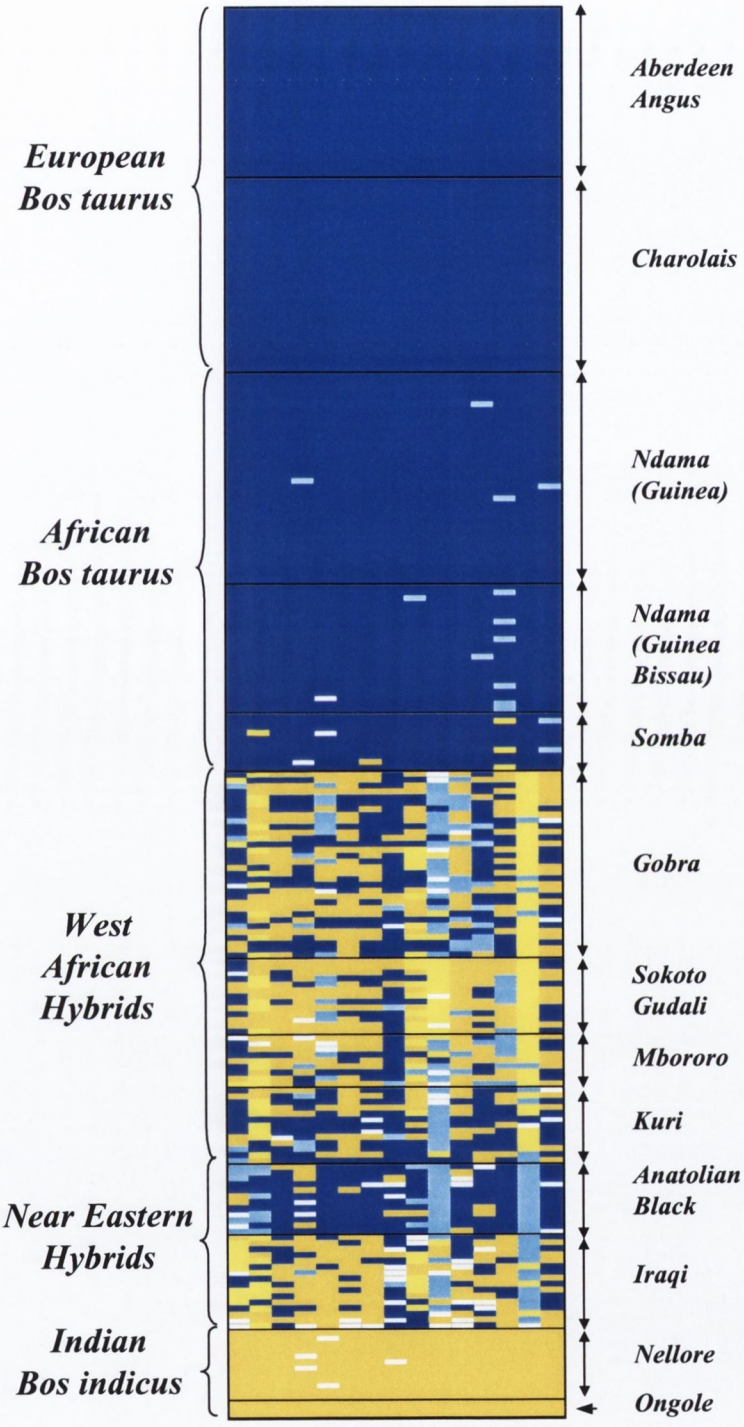


Figure 4.13 *X-chromosome haplotypes reconstructed from* STRUCTURE *allele designation. Each plot shows 240 horizontal lines, each indicating one individual. The breed to which with individual belongs is shown on the right hand side and by the dividing lines. The plots are vertically divided by the fifteen studied microsatellites in the order on which they are mapped to the X-chromosome. The different colours represent the ancestry status of the allele in the particular individual given its location. In the K=2 model plot the dark blue colour represents alleles that have been assigned to the ancestral population characterised by pure B.*

taurus with a probability of greater than 80%. The light blue colour shows those alleles that have been assigned to this population with a probability of greater than 50%. The dark yellow coloured rectangles show alleles that have been assigned with a probability of greater than 80% to the ancestral population characterised by pure B. indicus. Light yellow alleles have been assigned to this population with a probability of greater than 50%. In the plot that shows the results of the k-3 model the blue colours are now assigned to alleles characterised by the B. taurus population derived from the Near East and Europe, while the dark and light green rectangles highlight alleles that have been assigned as having ancestry derived from the population characterised by African B. taurus with probabilities of greater than 80% and greater than 50% respectively.

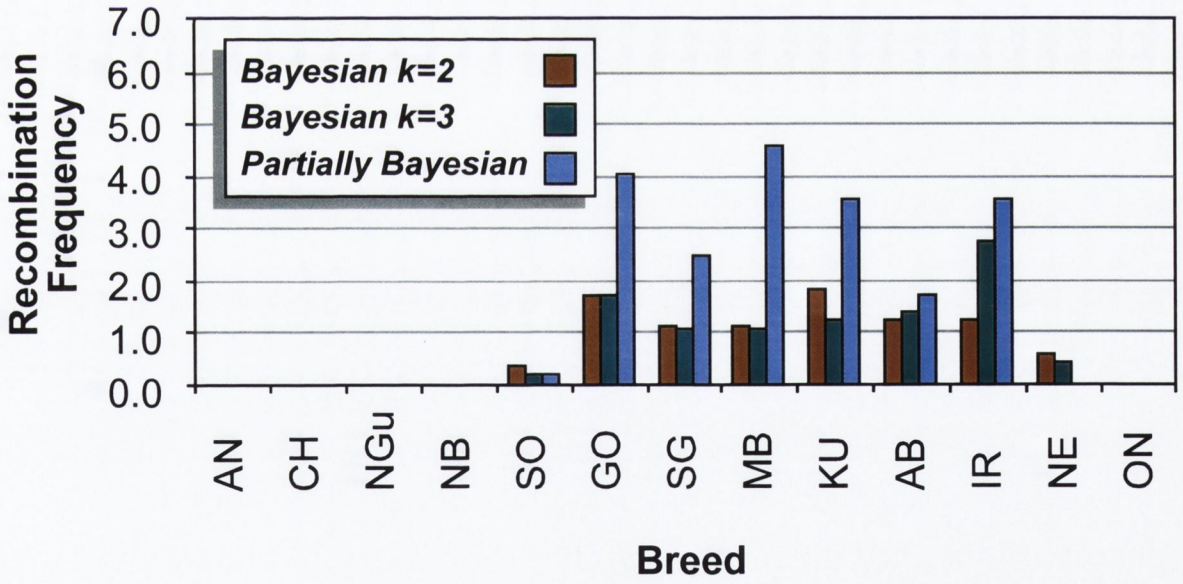
Figure 4.14 *X-chromosome haplotypes reconstructed from a partially Bayesian analysis of allele ancestry based on pre-assigned parental populations and breed admixture coefficients. The colours represent the same designations as in the k=2 model shown in* ***Figure 4.13.***



4.3.1.h Estimation of number of recombinations

The detectable number of recombinations between ancestral haplotypes was averaged within each breed and is shown in **Figure 4.15**. Alleles that were not assigned with a likelihood of at least 50% were excluded. Admixed breeds have significantly more recombinations within the surveyed regions than non-admixed breeds. More recombinations are detectable by the partially Bayesian analysis than the fully Bayesian analysis, as neighbouring allele ancestry is not used as a prior in the former. However, pairwise linear regression demonstrates that the estimated numbers of recombinations correspond well ($p<0.001$) between the different models.

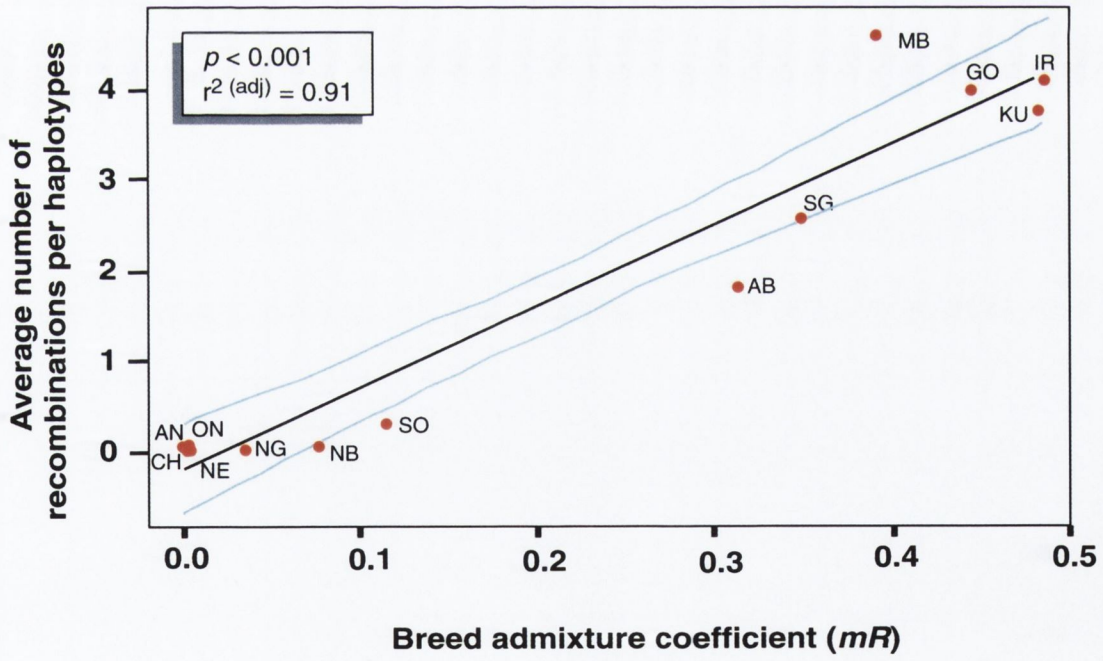
Figure 4.15 *Histogram of average number of recombinations in the surveyed region of the X-chromosome based on probable allele ancestry as defined by the three models listed in the figure legend. The least conservative method, the partially Bayesian, implies the greatest number of crossovers.*



The number of recombinations is related to the admixture history of the breed. While this is not the only factor that will influence the number of detectable crossovers, it seems intuitive that breeds with lower levels of admixture will have less recombinations that those with intermediate or high levels of admixture. A regression plot of average number

of crossovers and admixture coefficients (*mR*) is shown in **Figure 4.16**. If *mR* is greater than 0.5, 1-*mR* was used. The plot shows the high correlation and significant relationship between the 2 variables ($p<0.001$, r^2 (adj) = 0.91). If breeds have as many crossovers as expected, given their admixture level, they are likely to fall on or near the regression line. Breeds that are positioned off the line may have undergone less or more recombination than predicted given their admixture level. Breeds that are positioned below the line have experienced less detectable recombination than expected, while those above the line have undergone more than expected.

Figure 4.16 *Linear regression plot showing average number of recombinations detected by partially Bayesian methods plotted against admixture coefficient (mR). If mR>0.5, 1-mR was used as a measure of admixture proportion. Blue lines represent 95% confidence intervals. Each red circle represents one breed.*



Most of the cattle breeds in the present data group appear to have as many recombinations as would be expected, given their admixture coefficient. Interestingly, Anatolian Black, one of the Near Eastern hybrid breeds may have slightly fewer detectable crossovers than would be expected given its admixture proportion. This could be indicative of older

admixture but the difference between this and the other breeds does not seem to be significant.

4.3.2 RESULTS 2

Dataset 2 – 10 markers in twenty-two breeds

The previous analyses were repeated for a dataset comprising of 346 individuals from the 22 breeds listed in **Table 4.1**. These individuals were genotyped for the 10 microsatellite markers marked with an asterisk in **Table 4.2**. Linkage disequilibrium analysis was carried out only on breeds that were not included in the LD analysis in Section 4.3.1.

4.3.2.a Allele frequencies

Allele frequencies of the 10 microsatellites that were typed have been outlined in **Appendix A5**.

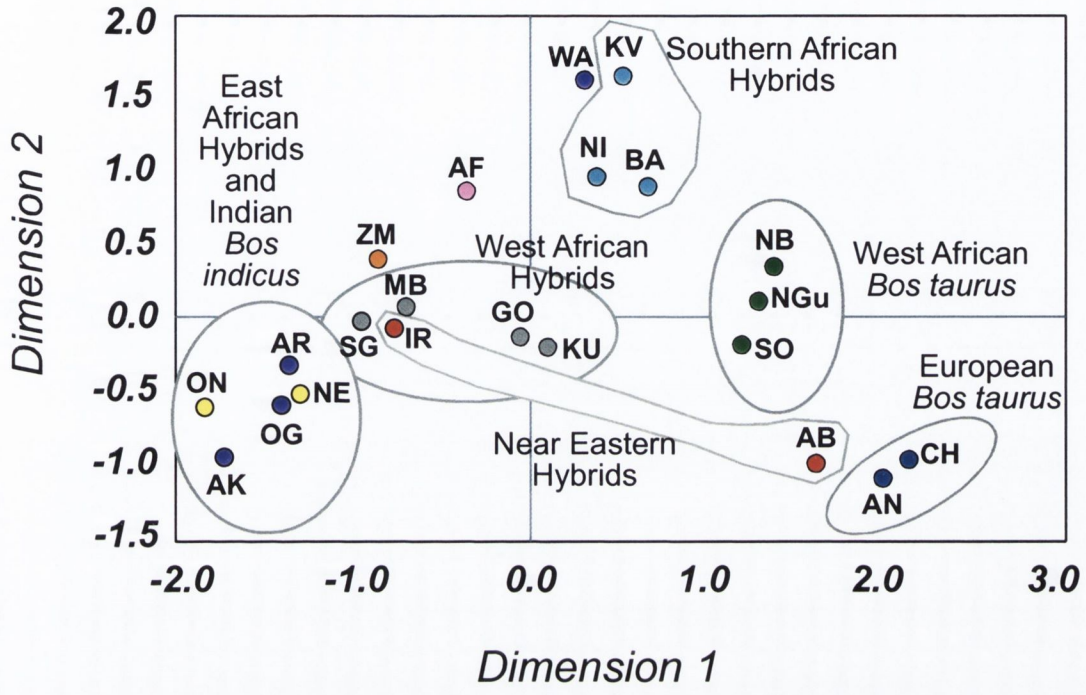
4.3.2.b Phylogeography

Phylogeographic analysis was repeated to establish the genetic relationships between all surveyed breeds, including both the Eastern and Southern African breeds and the previously analysed samples.

4.3.2.b.i Population Level

One again a 2-dimensional MDS analysis of population pairwise F_{ST} values is used to visualise the genetic distance relationships between the breeds. The plot (**Figure 4.17**) maintains that the main division is between *B. taurus* and *B. indicus* samples which are distributed by dimension one. Interestingly, the second axis of this scaling seems to divide European and Near Eastern taurine individuals from the Southern African breeds. The West African *B. taurus* breeds are found in an intermediate position.

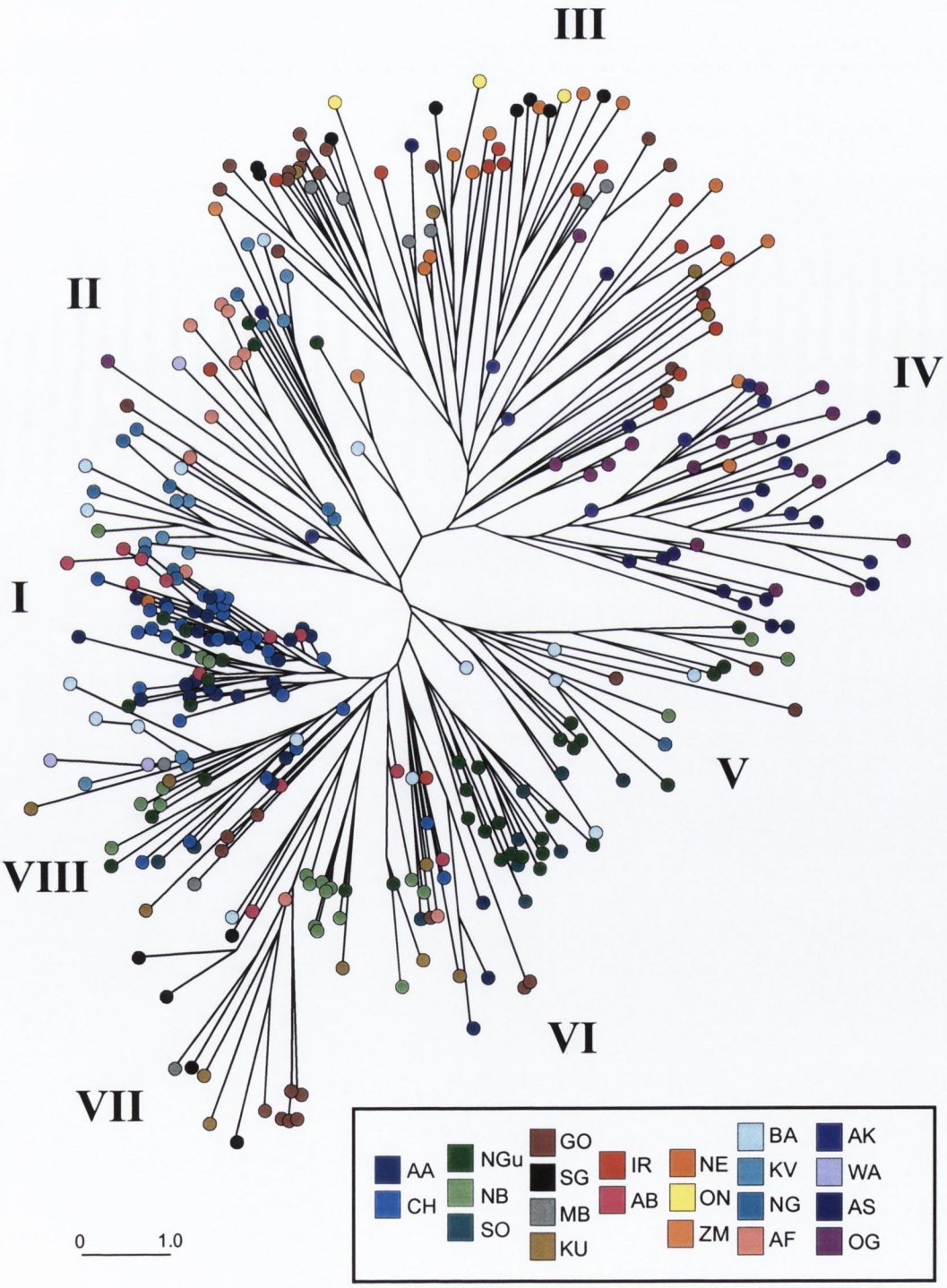
Figure 4.17 *Multidimensional Scaling plot of population pairwise F_{ST} values for 22 cattle breeds from all regional groupings, genotyped for 10 microsatellite markers. Breed codes are given in* ***Table 4.1.***



4.3.2.b.ii Individual Level

An unrooted neighbour-joining tree was constructed using allele sharing differences between individual chromosomes genotyped for 10 microsatelllite markers and is presented in **Figure 4.18**. The tree has several branches where the majority of individuals are from one of the regional groupings defined in **Table 4.1**. However, these 10 markers do not seem to resolve the groups as successfully as the 15 markers used for the analysis shown in **Figure 4.9**. While several clades contain mostly individuals from one of the regional groups (e.g. clades I, III, IV), there are also clades where individuals from different geographical locations group together.

Figure 4.18 *Unrooted neighbor-joining tree showing X-chromosome haplotypes constructed from 10 linked microsatellite markers. Clade I contains mostly European and West African Bos taurus individuals. Several Anatolian Black chromosomes also fall within this group. Clade II contains mixed African individuals, predominately the Southern and Western individuals with lower admixture levels. The large Clade III is populated mainly by Indian B. indicus individuals and African and Near Eastern individual with a large component of indicus admixture. Clade IV is made up of southern African samples and Clades V and VI are mostly Western and Southern African samples. Clade VII is a mixture of West African hybrid and non-hybrid individuals and Clade VIII seems to have individuals from all regions.*



4.3.2.c Mismatch analysis

Mismatch distributions were drawn using both the stepwise mutation model and the allele sharing methods to count the number of differences between individuals. These are shown in **Figure 4.19**. As was the case with **Figure 4.10**, the SMM distributions are ragged because of the highly polymorphic nature of the surveyed markers. However, the two models again produce very similar overall patterns. Once more, the West African and Near Eastern hybrid populations have the highest modes. Surprisingly, the Eastern and Southern African hybrids have modes that are almost identical to the European breeds. All the curves appear to be unimodal except for the Zebu Malagasy breed. This distribution looks bimodal, but is only constructed from 4 individuals so cannot be strongly interpreted.

4.3.2.d Linkage disequilibrium

LD was estimated for the breeds that were not included in the previous analysis (Section 4.3.2). The X-chromosome microsatellite data was combined with genotype data from fifteen unlinked markers and differences in levels of LD between linked and non-linked markers were assessed. The Zebu Malagasy breed was excluded from this analysis as many of the markers were monomorphic, probably as a results of the small sample size ($n = 4$). The results are presented in **Figure 4.20**. High levels of LD are detected in both in East African hybrid sample and the Afrikaner breed, especially between the linked markers. The East African hybrids have considerably higher levels of LD between non-linked markers. The recent history of the Afrikaner breed meant that it was appropriate to examine this group separately (CLUTTON-BROCK 1989). It is an artificially created breed and displays especially elevated levels of highly significant LD between linked markers. There are less significant pairwise associations between linked and unlinked loci in the Southern African hybrids.

Figure 4.19 *Mismatch distributions constructed from number of differences between genotypes constructed from 10 microsatellite markers. The Y-axis represents the number of individuals separated by the number of differences specified by the X-axis. The legend abbreviations are as follows: EUR=European B. taurus, WAT=West African B. taurus, WAH=West African hybrids, NEH=Near Eastern Hybrids, ZEB=Indian B. indicus, EAH=East African Hybrids, SAH=Southern African Hybrids, ZM=Zebu Malagasy and AF=Afrikaner.*

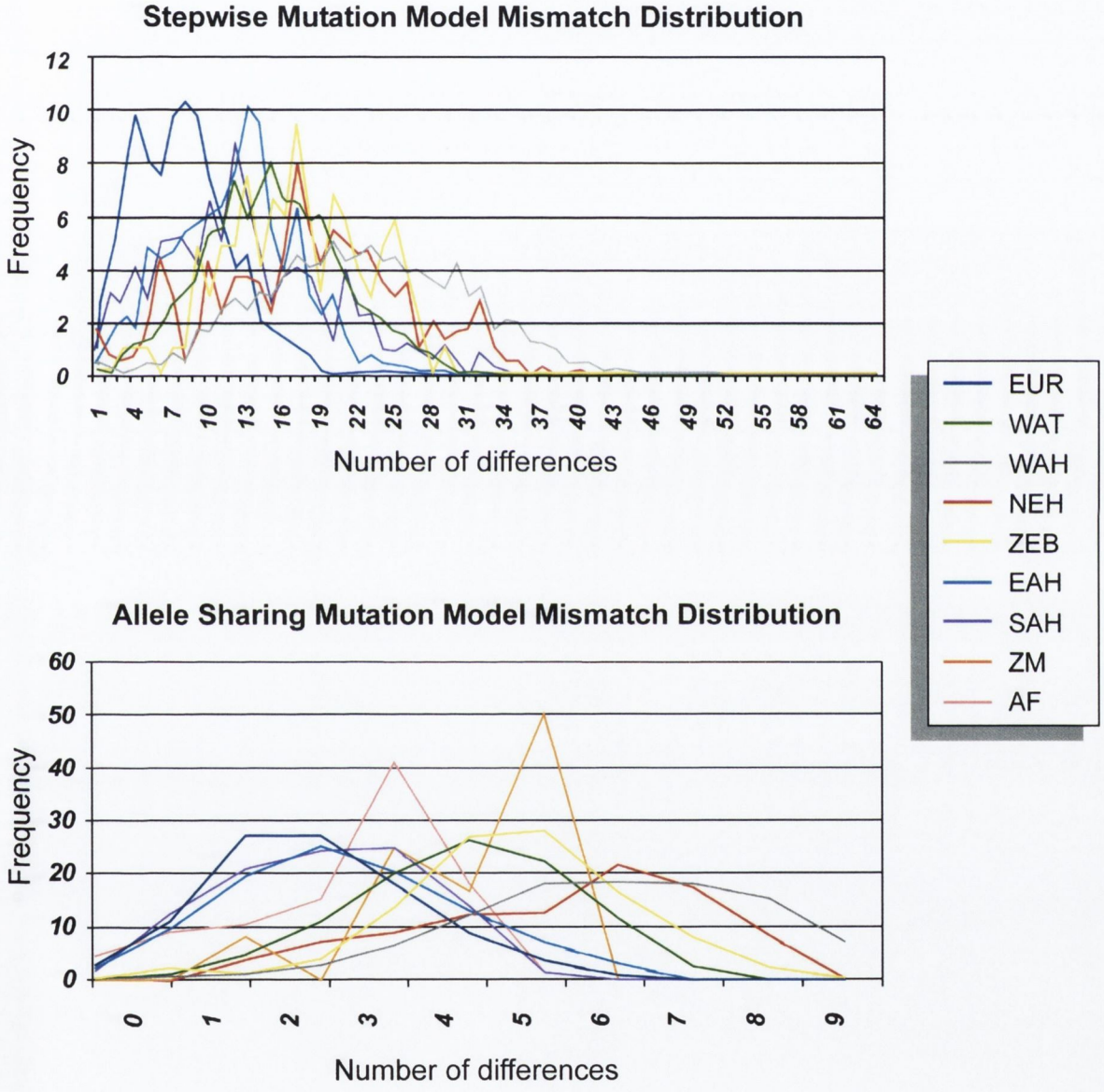
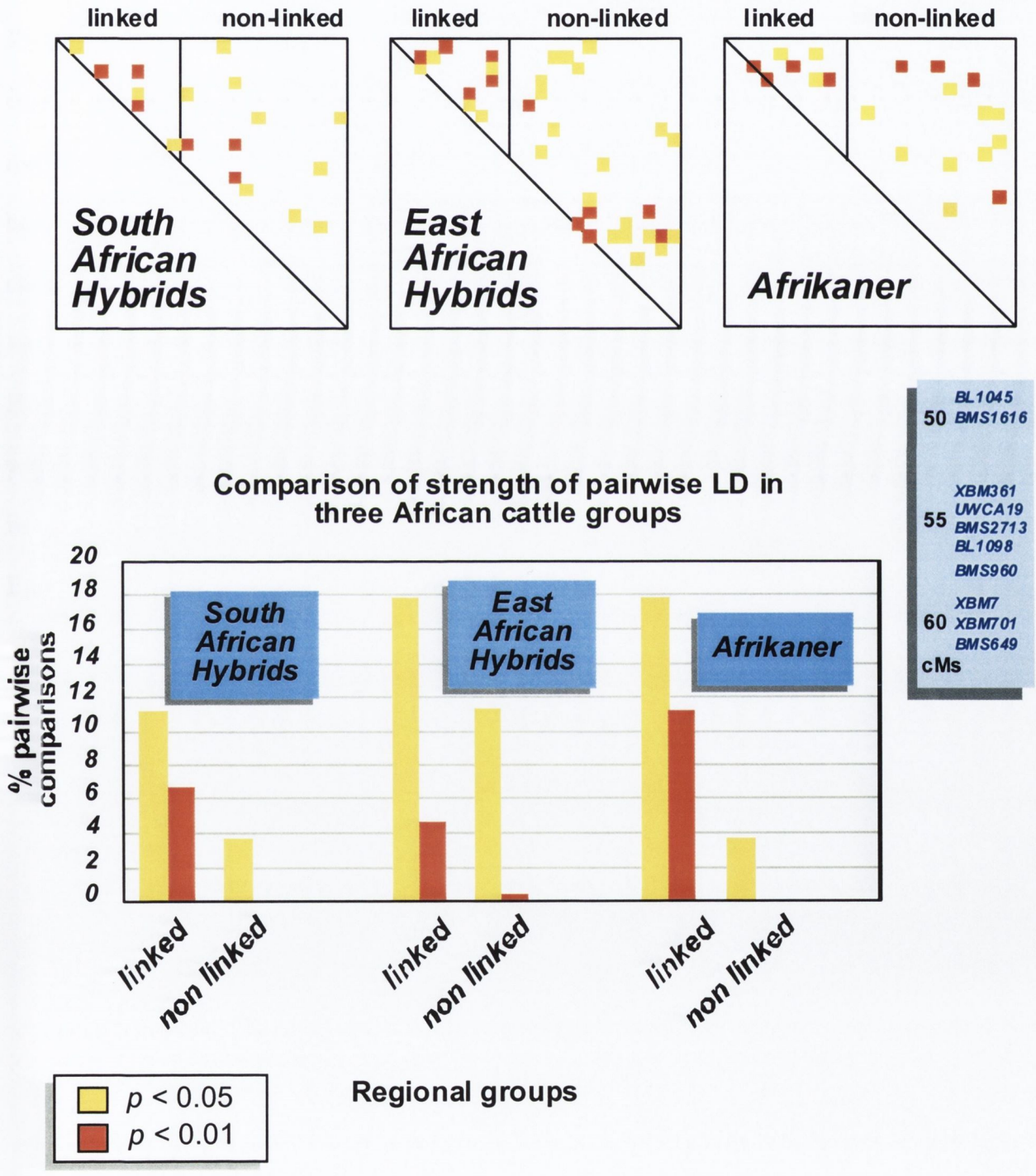


Figure 4.20 *Linkage disequilibrium results for Southern and Eastern cattle populations genotyped with 10 linked and 15 unlinked microsatellite markers. The matrix plots are reduced versions of the type of plot shown in Figure 4.10. They illustrate pairwise associations between markers at significance levels of p<0.05 (yellow) and p<0.01 (red). The markers run from left to right along the top of the plot (2-25) and from top to bottom along the left-hand side of the plot (1-24). The linked X-chromosome markers are ordered according to their mapped location as shown on the right hand scale. The histograms show the number of differences at the different significance levels for the linked and non-linked markers.*



4.3.2.e Fully and Partially Bayesian Analysis of Allele Ancestry.

Both the fully and partially Bayesian analysis of allele ancestry that were described previously were repeated for the twenty two breeds listed in **Table 4.1** which had been genotyped for the ten microsatellites marked with an asterisk in **Table 4.2**.

The results of the Bayesian approach are shown in **Figure 4.21**. As with the plots in **Figures 4.13 and 4.14**, the markers are ordered according to their mapped order on the X-chromosome. The k=2 model plot shows a clear separation between *B. taurus* and *B. indicus* breeds. European and West African taurine breeds and Southern African hybrid breeds are estimated to have predominantly *B. taurus* ancestry in this region of the X-chromosome. Anatolian Black again looks more similar to the European breeds than to the Iraqi population, the other Near Eastern breed included in the sample. The Western and Eastern African hybrids, Iraqi, Afrikaner and Zebu Malagasy breeds show mixed ancestry, with a higher proportion of *B. indicus* ancestry in the latter four. The k=3 model is shown in the right hand plot in **Figure 4.21** and clearly differentiates between African and European *B. taurus*. **Figure 4.22** presents the partially Bayesian results.

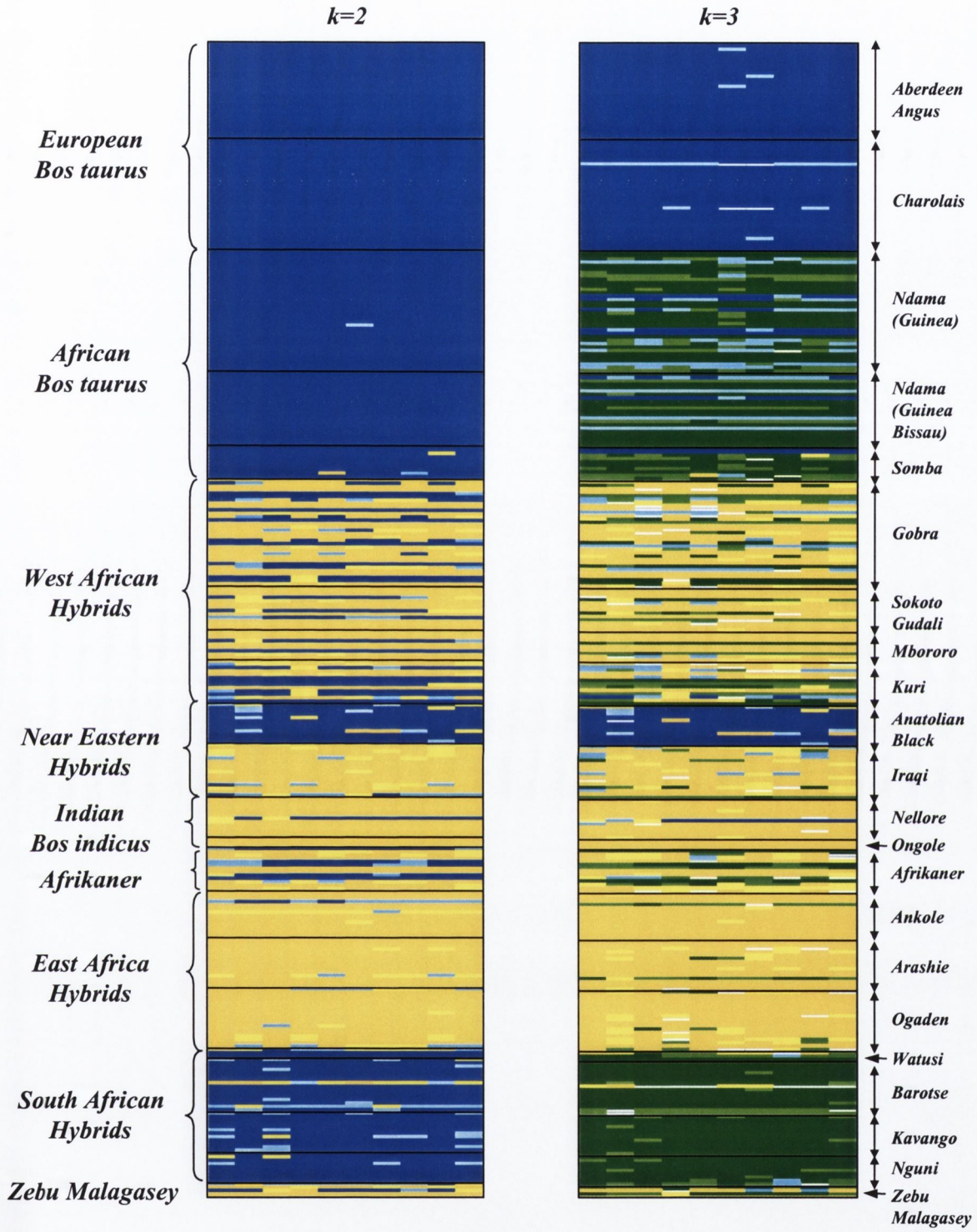
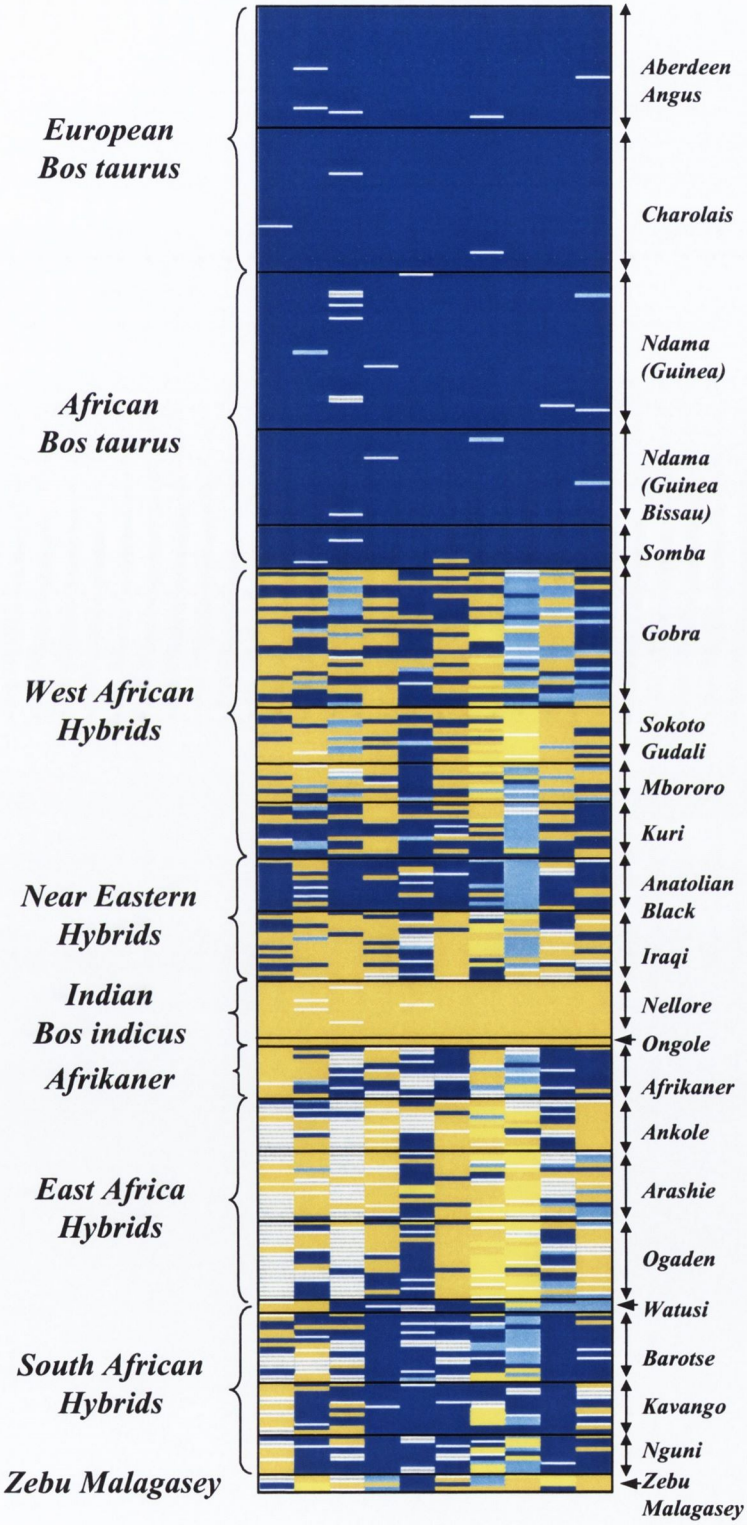


Figure 4.21 *Bayesian analysis of allele ancestry for dataset 2. Each plot shows 346 horizontal lines, each representing one individual. The breed to which with individual belongs is shown on the right hand side and by the dividing lines. The plots are vertically divided by the ten studied microsatellites in the order on which they are mapped to the X-chromosome. The different colours represent the ancestry status of the allele in the particular individual given its location. In the k=2 model plot the dark blue colour represents alleles that*

have been assigned to the ancestral population characterised by pure B. taurus with a probability of greater than 80%. The light blue colour shows those alleles that have been assigned to this population with a probability of greater than 50%. The dark yellow coloured rectangles show alleles that have been assigned with a probability of greater than 80% to the ancestral population characterised by pure B. indicus. Light yellow alleles have been assigned to this population with a probability of greater than 50%. In the plot that shows the results of the k-3 model the blue colours are now assigned to alleles characterised by the B. taurus population derived from the Near East and Europe, while the dark and light green rectangles highlight alleles that have been assigned as having ancestry derived from the population characterised by African B. taurus with probabilities of greater than 80% and greater than 50% respectively.

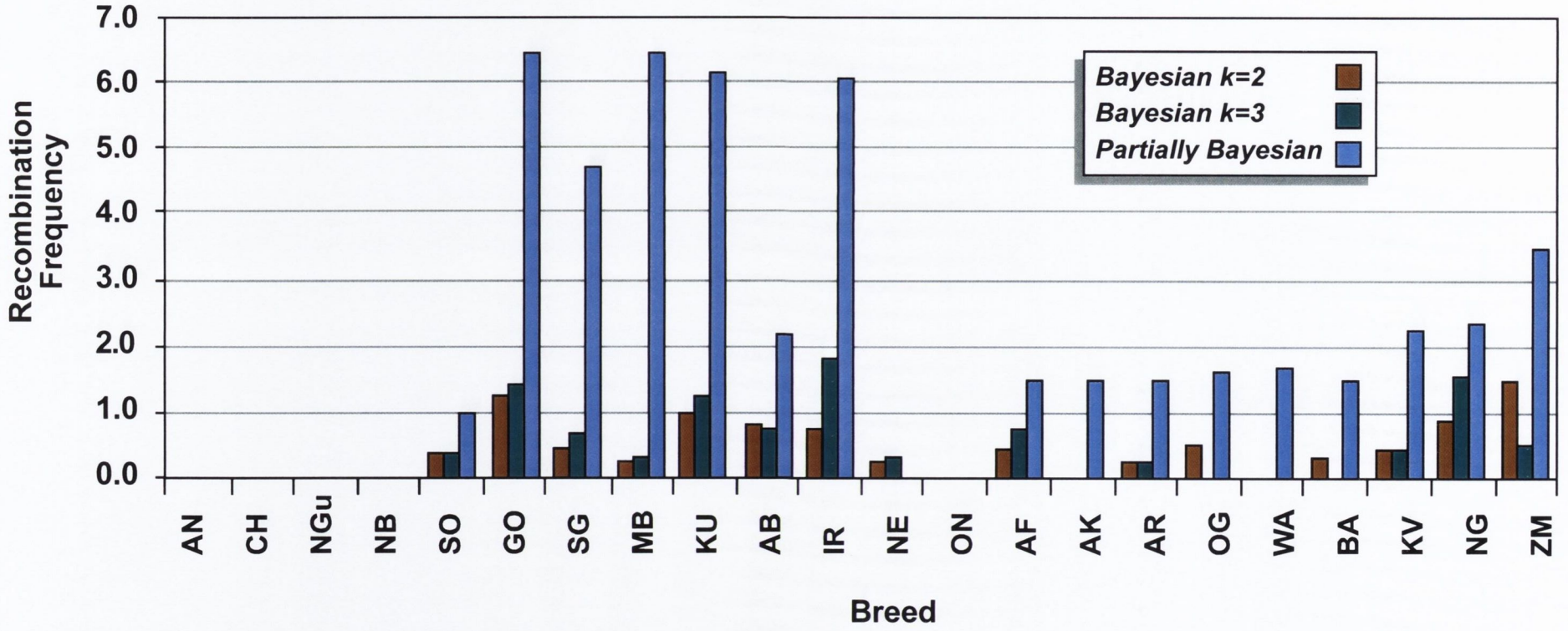
Figure 4.22 *Partial Bayesian allele ancestry assignment for ten microsatellite markers according to their mapped order on the X-chromosome. As a pooled sample of European and West African taurine was used as one of the representative parental populations, there are a number of alleles in the southern African breeds that cannot be assigned to either parental with a probability of greater than 50%. To avoid bias, these alleles were not included in recombination counting.*



4.3.2.f Estimation of number of recombinations

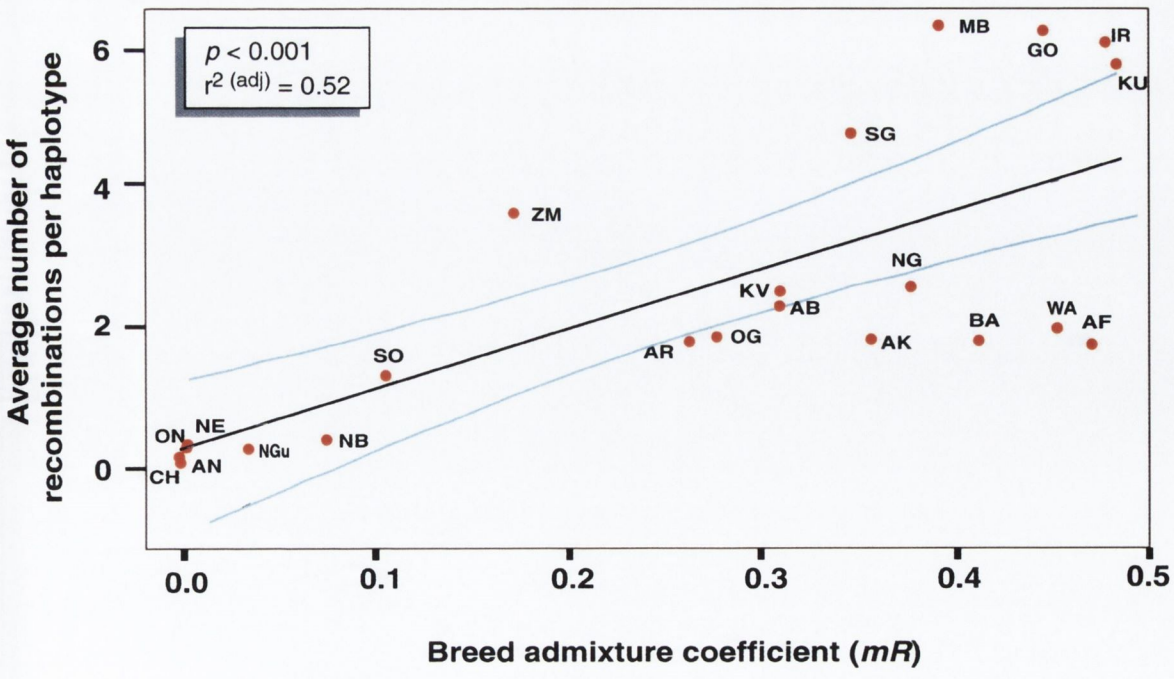
As in Section 4.3.1.h, recombinations per haplotype were counted under each of the three allele ancestry assignment models. The histogram of results is shown in **Figure 4.23**. The partially Bayesian model detects far more crossovers than the more conservative fully Bayesian analysis, but once again there is good proportional congruence between the different methods of recombination detection ($p<0.001$). Furthermore, the number of crossovers in hybrid breeds is far greater than in those populations lacking historical admixture.

Figure 4.23 *Histogram of recombination frequency in the surveyed region (ten microsatellites) of the X-chromosome based on probable allele ancestry as defined by the three models listed in the figure legend. As when fifteen markers are used, (**Figure 4.15**) the partially Bayesian approach is the least conservative and implies the greatest number of recombinations. The partially Bayesian approach does not detect any recombinations in the putative parental populations (AN, CH, NGu, NB, NE, ON) as this model presumes that they have admixture coefficients of 0.*



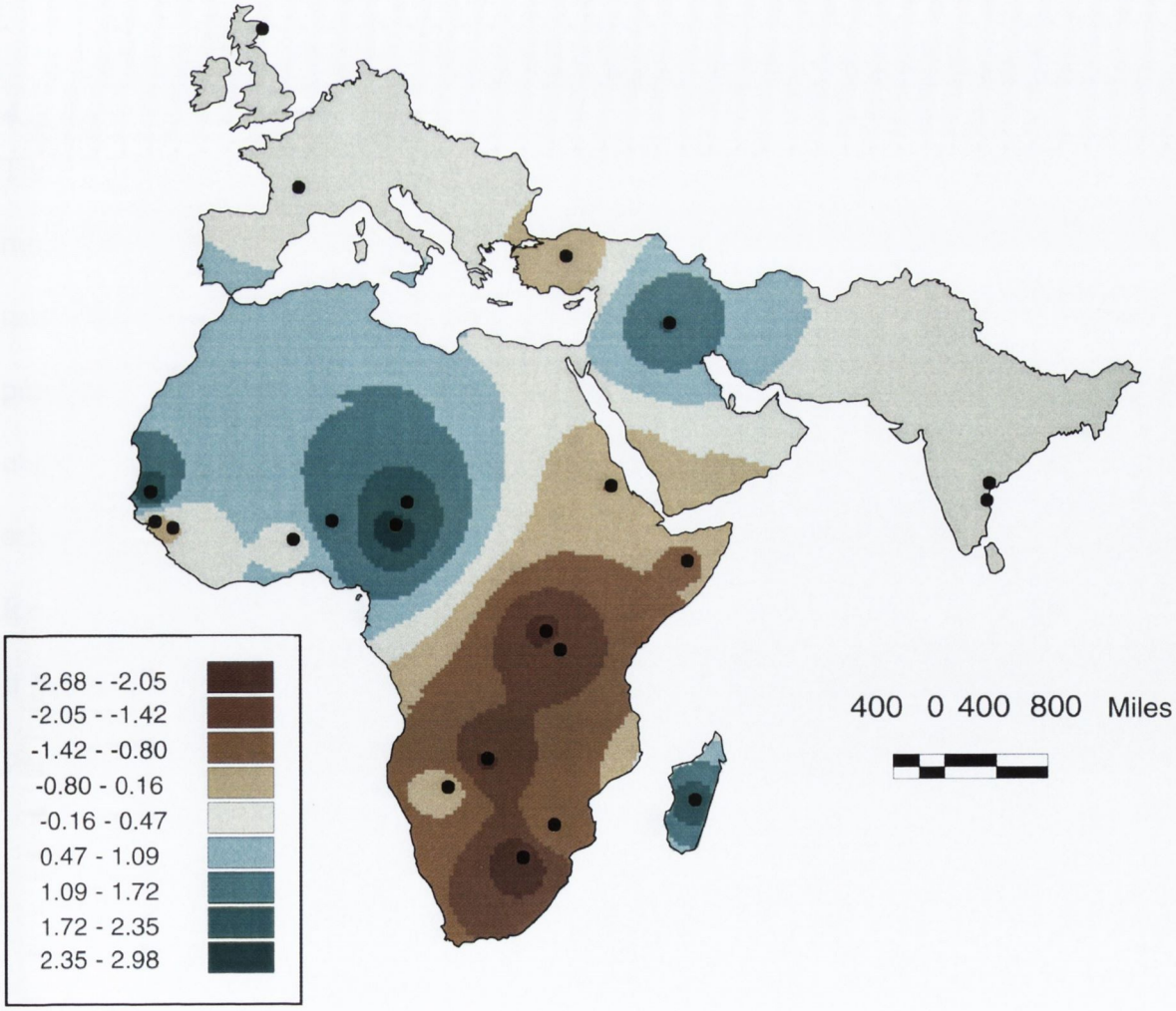
A regression plot of average number of crossovers and admixture coefficients (*mR*) is shown in **Figure 4.23**. This enables examination of whether a breed has experienced less or more recombination in this chromosomal region, given its admixture coefficient. The plot shows the correlation and highly significant relationship between the two variables ($p<0.001$, $r^2(adj) = 0.52$). There is more noise around the line than was observed for the previous dataset, which implies that breeds that were found on either side of the line might have different histories of admixture as they seem to have very different numbers of recombinations, even when admixture coefficients are quite similar (e.g. Mbororo and Barotse). Some geographical grouping of outliers is possible, as the West African hybrid breeds are concentrated above the regression line, while a group of southeastern African breeds are outliers below the line.

Figure 4.23 *Scatter plot of regression between admixture coeffiecient (mR) and average number of recombinations detected by partially Bayesian allele ancestry model. If mR>0.5, 1-mR was used as a measure of admixture proportion. Each red circle represents 1 of the breeds in the study. Blue lines represent 95% prediction confidence intervals.*



Residuals from the line were plotted on a map to facilitate distinguishing between the breeds that have different amounts of detectable recombination. This plot can be seen in **Figure 4.24**. Zebu Malagasy and some of the West African hybrid breeds (SG, MB, GO) are found above the line and are shown by the green contours. Iraqi also groups near these breeds. These breeds show markedly more recombination than expected under the linear model. Brown contours highlight the breeds that are found below the regression line and which have haplotypes showing relatively less recombination.

Figure 4.24 *Map of residual values for 22 cattle breeds. There is a significant relationship between breed admixture coefficient and detectable recombination in the surveyed region of the X-chromosome (**Figure 4.23**). Residuals from the regression between these two variables are plotted here, thereby highlighting breeds that show less or more recombination than average for this dataset.*



4.3.3 RESULTS 3

SNP analysis

4.3.3.a Allele frequencies

SNPs represent the most abundant form of genetic variation therefore the discovery of several variants within the sequenced intronic region is likely (KONFORTOV *et al.* 1999; MARTH *et al.* 1999). Several SNPs were discovered in introns 4,5,7 and 8 of the bovine ectodysplasin gene (**Figure 4.6**). Three polymorphisms in intron 8 were genotyped in 240 individuals from 13 breeds using SSCP-PCR techniques. These were denoted EDIINT8A, EDIINT8B and EDIINT8B. The allele frequencies for these SNPs are shown in **Figure 4.25**. These novel SNPs showed some discrimination between the *B. taurus* and *B. indicus* taxa.

4.3.3.b Allele ancestry

The data from the three SNPs were combined with genotypes from the closely linked microsatellite ED1MS1, and both the Bayesian and partially Bayesian analyses which were carried out on the microsatellite datasets were repeated for these 4 markers, inferring two parental populations. The results (**Figure 4.25**) show a broadly similar pattern to that observed for the previous two datasets. None of the chromosomes that have been schematically reconstructed from the fully Bayesian approach are fragmented. Recombinants are detected in the hybrid populations by the partially Bayesian analysis, but it is difficult to distinguish between crossover levels in different breeds. The markers are plotted in the order in which they are located along a 37kb sequence.

Figure 4.25 *Allele frequencies of novel SNPs discovered in intron 8 of the bovine ectodysplasin gene.*

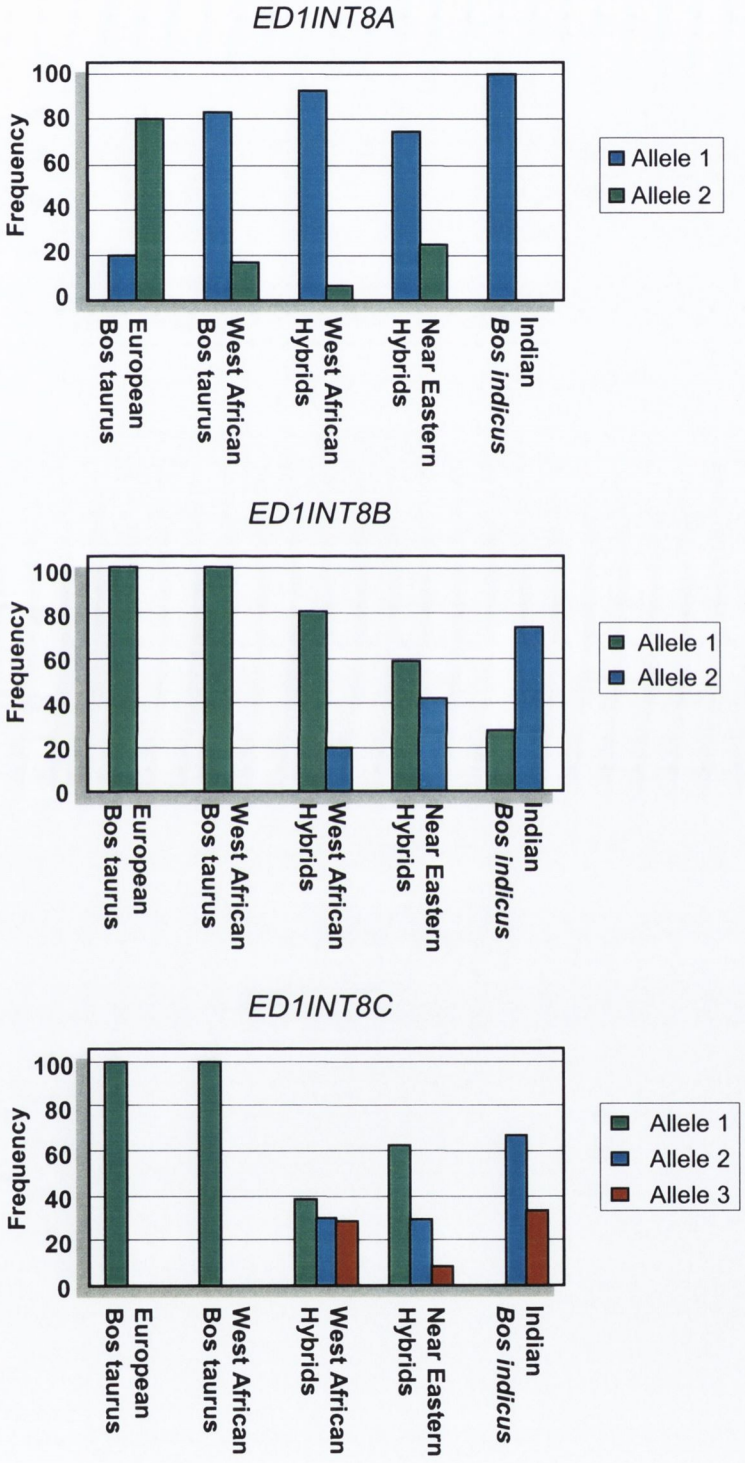
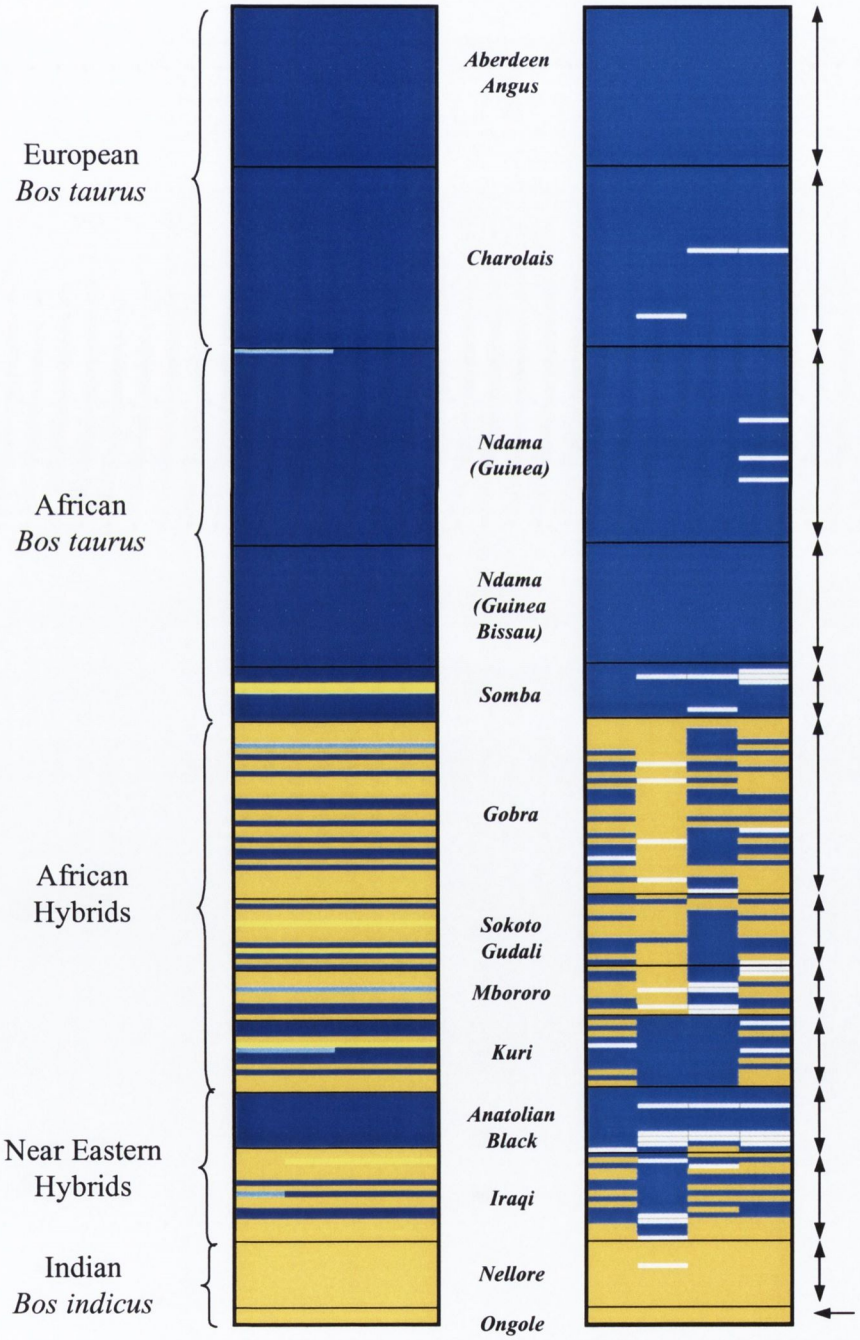


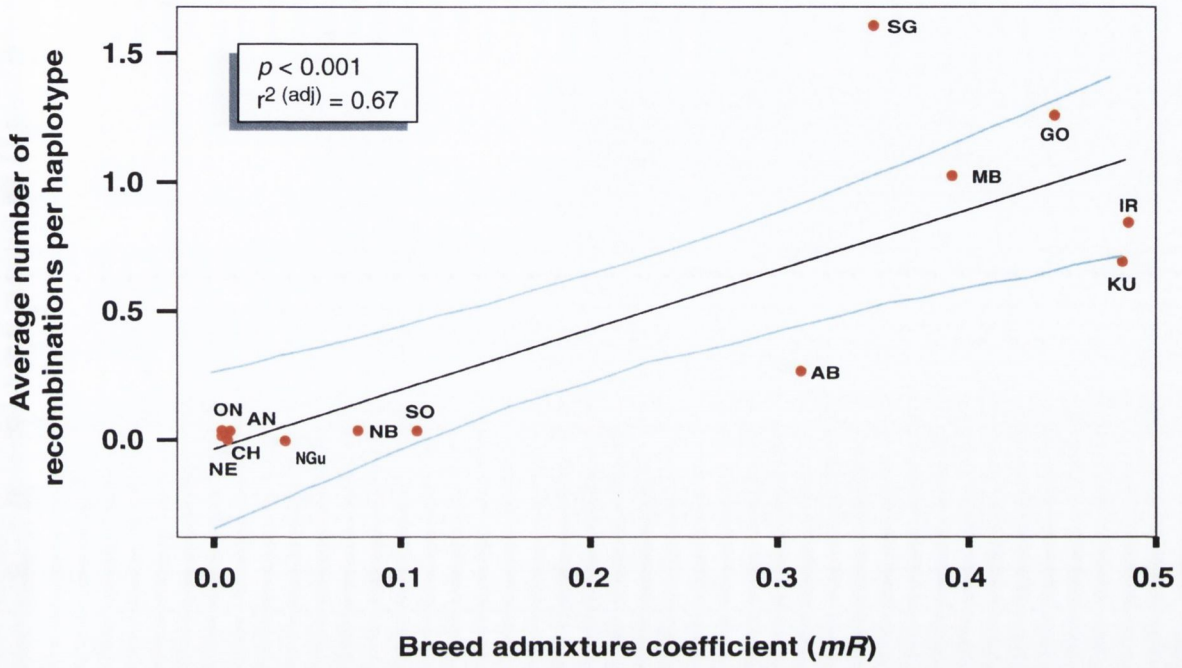
Figure 4.26 *Bayesian (left hand plot) and partially Bayesian (right hand plot) analysis of three X-chromosome SNPs and a closely linked microsatellite. The dark blue colour represents SNP alleles that were assigned to the B. taurus characterised population with a probability of at least 80%. Alleles assigned to this population with a probability of only 50% are shown by the light blue colour. The darker and lighter yellow colours show alleles that were assigned to the B. indicus characterised parental population with probabilities of 80% and 50% respectively. Non-assigned alleles are white. The plots are divided into 240 horizontal lines, each representing one individual. Horizontal black lines show divisions between breeds. The plots are divided into 4 vertical lines, the first represents the ED1MS1 microsatellite and the subsequent 3 are the ED1 gene intron 8 SNPS in the order to which they are mapped on the X-chromosome.*



4.3.3.c Estimation of number of recombinations

No recombinant haplotypes were detected using the fully Bayesian approach. This approach uses ancestry of surrounding markers and individual ancestry as priors in the estimation of allele ancestry. The method is conservative in terms of crossover analysis as it is more likely to infer that an allele shares the same ancestry as its neighbour. The partially Bayesian approach, which treats each marker independently, implies that some recombinations may have occurred (**Figure 4.25**). Direct counting and averaging the results for each breed estimated the number of putative crossovers. More crossovers were detected in the hybrid populations than in the populations with little of no history of admixture. There was a highly significant positive correlation between number of detected recombinations and admixture coefficients as defined by *mR* ($p < 0.001$, $r^2(\text{adj}) = 0.67$). A scatter plot for this regression analysis (**Figure 4.26**) shows that most breeds are found within 95% prediction intervals.

Figure 4.26 *Scatter plot of regression between admixture coeffiecient and average number of recombinations detected by partially Bayesian allele ancestry model between four physically linked X-chromosome markers. Admixture proportion was measured by mR, or 1-mR if mR>0.5. Each red circle represents one of the breeds in the study. Blue lines represent 95% prediction confidence intervals.*



Breeds that are positioned above the line may have more detectable recombinations that would be expected by chance. Sokoto Gudali (SG) is the only breed outside the 95% confidence intervals that appears to have a relative excess of recombination, a position it also occupies in **Figure 4.22**, but not in **Figure 4.10**. Those breeds that are found below the regression line have fewer recombinations than expected based on their admixture coefficients and include only Anatolian Black (AB), and the Kuri breed (KU). The result cannot be interpreted too strongly, as the number of markers is small. However, it is notable that this cassette of markers also highlights Anatolian Black as having a relatively low number of crossovers. This system may produce finer resolution as all markers are found within one contiguous 37kb database sequence, so their close proximity to each other and relative locations are certain.

4.4 DISCUSSION

The current study aimed to address several issues which are related both to assessing the potential for admixture mapping and to deciphering population histories. Firstly, it is essential to know over what chromosomal distances associations between markers descended directly from one parental population remain significant, and how this association can be accurately measured, and perhaps quantified. Secondly, determining the ancestry of relevant markers will be critical to deciding whether two markers have been co-inherited on a single chromosomal segment derived directly from one ancestral parental population. Finally the relationship between marker associations and the age, pattern and extent of admixture, can be deciphered by examining populations with different histories of introgression.

Cattle populations represent a promising model for the study of admixture-generated linkage disequilibrium and haplotype decay. Extant cattle have a complex evolutionary history that has begun to be elucidated over the last ten years using molecular techniques. While *B. taurus* and *B. indicus* can be interbred, they share a common ancestor at least 200,000 years BP and have unique genetic and morphological characteristics that have been combined in many different hybrid breeds (LOFTUS *et al.* 1994a; MACHUGH *et al.* 1997). These hybrids are very useful groups for genetic studies as there are extant taurine and zebu populations that are good representatives of the original parental populations that contributed to the modern day hybrids (MACHUGH *et al.* 1997).

Previous studies are supported by the phylogeographic analyses presented here. Multidimensional scaling analyses (**Figures 4.8 and 4.17**) of the cattle populations reveal a clear division between the taurine and indicine lineages as has been found with

mitochondrial DNA, Y chromosome and autosomal microsatellite systems (BRADLEY *et al.* 1994; LOFTUS *et al.* 1994a; MACHUGH *et al.* 1997). The difference between breeds from different parts of the African continent are also highlighted in **Figure 4.17**, where the separate axis of variance clearly positions West African taurine breeds as intermediates between Europe and South Africa.

This analysis provides confidence that the microsatellites used can discriminate between the different groups of cattle breeds that were selected and the phylogenetic trees of individual X-chromosome haplotypes, which are represented in **Figures 4.9 and 4.18**, provide an interesting comparison. While most haplotypes group with other haplotypes from the same region, some individuals are found within clades characterised by another regional group. These individuals probably represent introgressors, but could illustrate that over any genome region, an individual may derive most of their genome from one ancestor. This may not necessarily be reflective of the whole genome admixture proportion, a more accurate estimation of which will be provided by randomly selected markers.

The fact that all of the markers used were on the X-chromosome may also influence the phylogeographic results. Male-driven introgression has resulted in the vast majority of African cattle breeds having a zebu Y-chromosome, however they have retained taurine mitochondrial DNA (BRADLEY *et al.* 1998). The genetics of the X-chromosome are also different to the autosomes, as there is no opportunity for significant recombination in males therefore the effective population size is smaller (WILSON and GOLDSTEIN 2000).

Three different types of analyses were used to examine the pattern of haplotype decay in the bovine genome: mismatch analysis, examination of patterns of linkage disequilibrium and finally, likelihood of ancestry was estimated for each allele. All of the microsatellite markers that were used have been mapped to the X-chromosome and were typed only in males. This allowed the unambiguous assignment of a haplotype to an individual without any pedigree information and without using maximum likelihood or expectation-maximization algorithm methods.

4.4.1 MISMATCH DISTRIBUTION

It has been hypothesised (Section 4.1.3.a) that the mode of mismatch distributions in recently created hybrid populations may be higher than for non-introgressed or older hybrid populations (which have not undergone subsequent introgression). Moreover, as divergent haplotypes recombine over time the average distance between them decreases. No population in the current analysis had a bimodal curve, which could imply very recent admixture, but the relative positions of the unimodal curves suggest some differences. Mismatch analysis shows a clear differentiation between hybrid breeds in West Africa and the Near East, and populations from east and southern Africa (**Figure 4.19**). On comparison with the non-hybrid populations, there is more divergence between haplotypes in the former two groups, whether 10 or 15 markers are examined (**Figure 4.10**). By contrast, the latter groups are found to have fewer within-population differences. In fact, the modes of these distributions are almost identical to that estimated for the non-hybrid European population. The seeming homogeneity of these populations may be a signature of their more recent colonisation of this area, combined with modern incursion from *B. indicus*.

Laan and Paabo (1997) found that the Saami, a human population with high levels of LD extending over 5.6-5.8 cM, had the highest haplotype mismatch variance of a group of Scandinavian populations. This is characterised by groups of related haplotypes that differ substantially from one another. The hybrid populations in the current study do not display significantly different levels of mismatch variance when compared to non-introgressed populations (LAAN and PAABO 1997).

4.4.2 LINKAGE DISEQUILIBRIUM

Sherrington and colleagues used microsatellites to detect linkage disequilibrium in 1991 and showed that they were stable through evolution and in the germline in humans (SHERRINGTON *et al.* 1991). In simulations it was found that the more highly polymorphic loci are better at detecting significant non-random association (SLATKIN 1994), and microsatellites perform better than SNPs (VARILO *et al.* 2003). While both SNPs and microsatellites are used for association mapping in humans, it has been suggested that microsatellites could be used as mapping tools in livestock, as a lower density will be required (FARNIR *et al.* 2000). In the present study the markers selected displayed a range of polymorphism. The less variable ones may be more suitable as ancestry-diagnostic markers but highly polymorphic microsatellites might be more suitable for detecting linkage disequilibrium. There is a large body of published work that demonstrates that linkage disequilibrium can extend over large regions of the genome many generations after admixture, especially if there is continuous gene flow (FARNIR *et al.* 2000; KAPLAN *et al.* 1998; LAUTENBERGER *et al.* 2000; PFAFF *et al.* 2001). Admixture level and the time at which a hybridisation event occurred are closely related to LD patterns (CHAKRABORTY and WEISS 1988).

Extensive LD was found in this region of the X-chromosome in some cattle populations of mixed origin, especially those *B. taurus*/*B. indicus* hybrids that are found around the tsetse zone of West Africa (**Figure 4.11**). This meta-population exhibits LD, which is significant at a *p*-value of 0.05, over the entire surveyed region. Most of the other regional populations show some associations between linked microsatellites at this significance level. The Near Eastern population showed intermediate levels of LD which seemed to be more localised on the chromosome and only extended over 10 cM at a significance level of p=0.05 (**Figure 4.11**). Both the East African hybrid breeds and the Afrikaner breed of

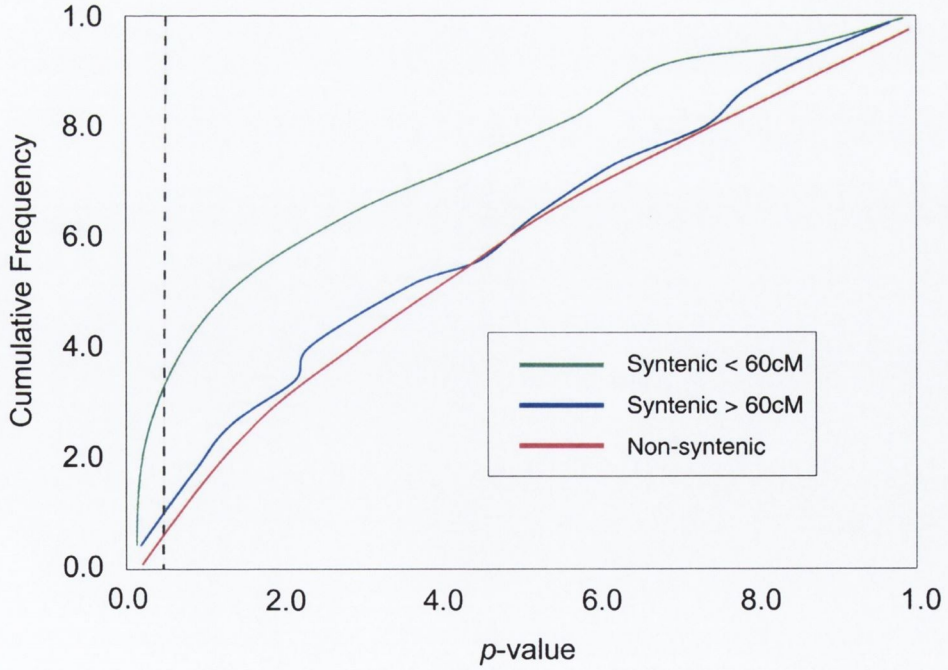
South Africa have similar intermediate levels of LD (**Figure 4.20**). An increase in the stringency of the association test to p=0.01 (**Figures 4.12 and 4.20**), or even p=0.001 (**Figure 4.11**), caused most of the disequilibria to cease to be significant, except the West African hybrid population.

Previous studies on cattle have found substantial LD, but have been conducted only in economically important European breeds. Farnir and colleagues (2000), who surveyed 284 autosomal markers in Holstein-Friesian cattle populations, found that intra-chromosomal LD extends over several tens of centimorgans and is also common between non-syntenic loci. These high levels of overall LD may be a result of a small effective population size (LONJOU *et al.* 1999), which could be as low as 50 for the breed which comprises 1.2 million lactating cows, or admixture (FARNIR *et al.* 2000).

A second study utilised two groups of microsatellite markers from chromosomes 2 and 6, which were genotyped in 50 young dairy bulls. The two groups were comprised of six markers located over 38 cM and seven markers over 20 cM respectively. LD that was significant at p=0.0007 was detected only between pairs of loci within the syntenic groups, extending up to 10cM. No LD was detected between markers on different chromosomes. This lack of significant LD between linkage groups led the authors to conclude that association studies in cattle could be more successful and cost effective than traditional quantitative trait loci mapping techniques, which rely on more extensive LD between trait and marker loci in family-based samples than in an out-bred populations sample (TENESA *et al.* 2003).

Sheep are the only other livestock breed for which there is published information about LD levels (KALINOWSKI and HEDRICK 2001; MCRAE *et al.* 2002). McRae and colleagues assessed LD in two New Zealand sheep breeds: Coopworth and Romney. Coopworth is a hybrid breed that was developed from Border Leicester and Romney about 10 generations ago, though selection is a facet of the history of both breeds. As in dairy cattle, LD that extended over tens of centimorgans was observed. The researchers found that LD decreased with distance and was observed at a greater frequency for syntenic markers separated by less than 60cM on all 10 chromosomes, assessed for 90 markers, in the Coopworth samples (**Figure 4.27**).

Figure 4.27 *(reproduced from MCRAE et al. 2002) Cumulative frequency plots of the significance of LD between markers genotyped in an admixed sheep population. Marker pairs are classified as nonsyntenic, syntenic but separated by >60cM, or syntenic but separated by <60cM. The dashed line shows p=0.05 and demonstrates that syntenic loci which are linked <60cM show significantly more LD that non-linked loci.*



The Romney breed was found to have lower overall levels of LD and had no significant associations between markers that were separated by distances greater than 60cM or non-syntenic markers when genotyped for 26 markers on one chromosome. However, no decline in LD with distance was observed for markers spaced at distances greater than 10 centimorgans. The authors noted that non-admixed breeds should be more appropriate for fine mapping. However, admixed breeds are likely to be useful for broader gene localisation, with the caveat that such studies would have to be carefully controlled for false positives resulting from high overall LD levels (McRae *et al.* 2002).

4.4.2.a Linkage disequilibrium and mapping

Interpretation of the present LD results is greatly facilitated by a comparison between the background level of LD in the entire genome, and the level that is detectable between the X-chromosome linked markers. Successful gene mapping studies that exploit linkage disequilibrium between neutral markers and phenotypes and/or candidate genes will require accurate information about background levels of LD (Carlson *et al.* 2003; Rannala 2001; Rannala and Slatkin 2000). Histograms comparing the percentage of possible comparisons between linked and unlinked markers at two significance levels confirm the status of the West African hybrids as the populations with the highest LD levels between linked markers, even when this level is compared to the background level for unlinked markers (**Figures 4.12 and 4.20**). The dataset incorporating 10 microsatellites reveals relatively high levels of LD in the east African hybrid breeds and the Afrikaner breed. As might be expected, most of the linked markers that are highly significantly associated in the West African hybrid populations are mapped beside each other. Reciprocal recombination has been found to be the major mechanism of haplotype destruction (Wiehe *et al.* 2000).

It is not uncommon to find significant LD over considerable regions of the X-chromosome in admixed populations. Wilson and Goldstein (2000) carried out a study of X-chromosome linkage disequilibrium in the Lemba, an admixed human population who are hypothesised to have shared ancestry with two modern groups: the Ashkanazi Jews and the Bantu. They found much higher levels of LD in the hybrid population that extended up to 21 cM. However, they also found LD in 25-38% of marker pairs separated by 2 cM in populations without significant evidence of admixture. As some of the markers in the current study are very closely linked it is not surprising that LD is seen in all of the populations to some extent. This might be expected as a result of the microsatellites close proximity on the X-chromosome and the possibility of limited admixture in all populations. However, the stringent *p*-value of 0.001 shows definitively that the West African hybrid population has significantly more LD than the others.

It seems unlikely that LD is a simple function of distance between markers. Dramatic heterogeneity of LD levels across a region of the X-chromosome has been observed in both out-bred and isolated European human populations (TAILLON-MILLER *et al.* 2000). However, other studies have reported far greater uniformity, including one that investigated four populations of Europe derivation: Afrikaners, Ashkenazim, Finns, and East Anglian British. This study found a similar inverse relationship between LD and distance in each of three genomic regions in each population (DUNNING *et al.* 2000). An extensive survey of 127 polymorphisms in three genomic regions in a sample of individuals with British ancestry found that approximately 45% of LD was related to position effects between markers (ABECASIS *et al.* 2001).

The ability to distinguish between the background level of disequilibrium and LD that is detectable between physically linked markers by the strength of association between markers is critical, if association between markers is to be used in mapping. False positive results in such studies are a genuine impediment to successfully locating genes of interest. The results in the current study suggest that avoidance of false positives in gene mapping studies is best achieved by considering only candidates that are in highly significant disequilibrium (e.g. $p<0.01$), and which are associated with more than one neutral marker (GARNER and SLATKIN 2003). If this more stringent cut-off value was applied to the data presented by McRae and colleagues (2002) and reproduced in **Figure 4.27**, the number of type II errors would be greatly reduced. Background associations tend to occur in isolation and the majority cease to be significant when a stringent p-value is investigated. The potential to use LD to map genes in cattle populations is excellent, and has been exploited, in combination with linkage methods, to fine map a gene involved in milk fat content (FARNIR *et al.* 2002).

In summary, it is possible that mapping using LD could be more successful in livestock than humans because many of the forces that produce LD, including small effective pop size, genetic drift, selection and admixture, are characteristic of domestic breeds (MCRAE *et al.* 2002). However, excessively high levels of LD mean that fine mapping is likely to be challenging, and that there is a greater chance of detecting associations between non-linked markers (FARNIR *et al.* 2002). Nonetheless, it is clear that the extent and homogeneity of LD in livestock needs to be established, especially in those breeds that are either economically important or possess genetic traits of interest, such as disease resistance.

4.4.2.b LD as a tool for evolutionary inference

LD is also a useful tool for historical of evolutionary inference (NORDBORG and TAVARE 2002). To determine the significance of the observed LD and assess its likely historical interpretation, an inspection of LD levels in other hybrid populations is informative. This high level of admixture-generated LD is not uncommon and may be due to the phenomenon of continued introgression from one or both of the parental populations. Lautenberger et al. (2000) found a highly significant LD score over 8 cM, which remained significant over 30 cM in African Americans. They calculated that only 69 generations would be required to halve the LD between markers 1 cM apart, but only 6.6 generations if 10 cM separated the markers and concluded that recurrent introgression must be the cause of these high levels of LD. As the generation time is cattle is much lower than in humans, it is likely that LD could be broken down quite rapidly in a panmitic population. As such the presence of such high levels of LD is indicative of recent or continual admixture. Although the manner in which admixture occurs will affect LD, it remains a useful measure to estimate an effective time since admixture and could be corrected for multiple waves of introgression.

The very high level of LD in some West African population may partially be explained by either recent or continuous gene flow (CGF) from zebu populations (**Chapter 1**). CGF reintroduces LD between loci in each generation. It has been previously documented that LD between neutral markers is caused in hybrid zones and that a barrier to gene flow (such as the tsetse zone) will also influence LD levels (BARTON and SHPAK 2000). LD levels at the centre of the hybrid zone are likely to be high and caused by an influx of parental combinations of genes into this region (BARTON and GALE 1993). This influence is likely to create a CGF situation in populations in the vicinity of a hybrid zone barrier.

The West African hybrid cattle populations in this study may have undergone successive waves of introgression, which contributed to their high levels of LD. Continuing introgression may be a factor. The relatively elevated levels of LD in the East African breeds could also be the result of recent or continuous admixture. Hanotte and colleagues (2002) have demonstrated that the majority of *B. indicus* arrived in east Africa before dispersing throughout the continent (HANOTTE *et al.* 2002).

It is now clear that there are both human and livestock populations with extensive LD rendering Kruglyak's early estimation (KRUGLYAK 1999b) that there would be no useful LD spanning more than 3kb in human populations untenable. Furthermore, cattle populations of mixed origins haven complex patterns of LD that are far ranging, and can be readily detected using polymorphic marker systems. This phenomenon may prove very useful in elucidating population histories and for the mapping of genes influencing both monogenic and complex traits. However, it is clear that understanding population history is also critical to successful mapping (VARILO *et al.* 2003).

4.4.3 ALLELE ANCESTRY AND RECOMBINATION

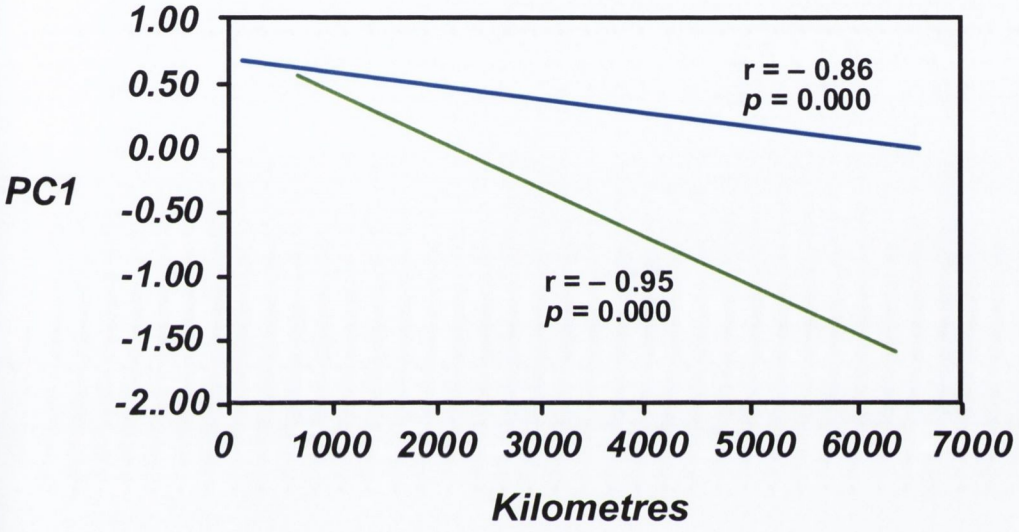
In the absence of mutation, every portion of a chromosome in a hybrid individual should be directly traceable to one of the individual's ancestors. When two distinct gene pools meet, a mosaic genome is produced (CHAPMAN and THOMPSON 2002). Intact segments from each contributing founder populations will be present in hybrid individuals (NORDBORG and TAVARE 2002). Genomes that have undergone recent hybridisation will have largely intact ancestral haplotypes, while recombination will act to disrupt these haplotypes in subsequent generations (WIEHE *et al.* 2000). The level of divergence between the contributing parental populations will determine whether it is possible to assign definite ancestry to an allele in a hybrid individual. The correct assignment of an allele in a hybrid population to a parental population is fundamental for the accurate study of admixture (MCKEIGUE 1998).

The current study employed two methods to assign ancestry to alleles, one partially Bayesian method and one fully Bayesian method. Each of these method estimates a probable ancestry for each allele in each X-chromosome haplotype. Haplotypes constructed from 15 microsatellites (**Figures 4.13 and 4.14**), 10 microsatellites (**Figures 4.21 and 4.22**) and one microsatellite and 3 SNPs (**Figure 4.26**) were analysed. The amount of fragmentation in haplotypes was assessed by estimating the number of detectable recombinations that would have occurred to produce the observed haplotype from intact *B. indicus* or *B. taurus* haplotypes. Given that this value is a function of breed composition, this value was compared to a breed admixture coefficient. There is a positive linear relationship between admixture coefficient (as represented by *mR*) and the degree of recombination within haplotypes. The partially Bayesian inferred ancestries (where each locus is counted independently) were used to estimate the number of recombinations.

Outliers to the correlation between these two measures warrant investigation. While none of the breeds that were genotyped for 15 markers showed a significant deviation from the linear pattern, this is not the case when a comparison is made between the larger suite of breeds that were genotyped for 10 markers. The residuals of the latter regression were plotted according to geographical location (**Figure 4.24**) and show marked regional distribution. The results suggest a dearth of detectable recombination in cattle populations from the east and south of the African continent. This finding supports a scenario of recent introgression of *B. indicus* material into these breeds, as the degree of recombination would be expected to accumulate with time. By contrast there seems to be an excess of detectable crossovers in breeds from the surround of the tsetse zone of West Africa, and in the Iraqi breed of the Near East. It is more difficult to resolve this finding with previous results, which show high levels of LD in this region (**Figure 4.12**), hypothesised to be the result of continual gene flow caused by a barrier to gene flow.

Hanotte and colleagues used principal components to examine the variance within Africa and have suggested that there are two different patterns of zebu introgression in the African continent that can be distinguished by the morphology of hybrid breeds (**Figure 4.28**). They have suggested that these two components could relate to different waves or rates of zebu incursion, a theory not without support in the archaeological literature (CLUTTON-BROCK 1989; MARSHALL 2000).

Figure 4.28 *(reproduced from HANOTTE et al. 2002) Regression of principal component 1 (PC1) of Hanotte et al, 2002, versus distance from the Horn of Africa (represented by the distance from the most northeasterly population sampled). PC1 was found to account for 38% of total allelic variation, the result of B. indicus admixture from a major secondary introduction to the continent. This analysis was conducted separately for populations (n=50) denominated as B. indicus (blue) and those denominated as B. taurus (green). There are two separate relations of this component, each of which strongly correlates with the level of zebu genetic background in today's African breeds. The authors hypothesize that this indicates the presence of two different patterns of zebu introgression.*



The indicine component of the *B. taurus* breed has been "bred in", whereas the other component is thought to have "walked in" more recently and is found in those breeds with a zebu morphology. However, there are breeds classified as both *B. taurus* and *B. indicus* type in both the south-eastern and western groups in the present study, so it is not possible to postulate that the signature of the two waves of zebu incursion is being detected here. The difference that is observed between the two areas may be the results of the different histories of cattle in the regions. The earliest cattle of Africa were found in the north of the continent. The major *B. indicus* introgression occurred in the East. Given proximity and antiquity, nomadic pastoralists have probably been co-herding animals from the two cattle lineages along the Sahel corridor for many thousands of years. The taurine component in

West African breeds may share a common ancestor with breeds like N'Dama, but is not necessarily derived from N'Dama.

On the other hand, cattle are more recent arrivals in the south of the continent, and while the first domesticated cattle to reach the south may have undergone limited zebu introgression before migrating towards the Cape, the majority of contact, especially with zebu breeds that have entered the continent by a maritime route has probably taken place in the last 1600 years (SMITH 2000). There is limited archaeological evidence for exactly when hybridisation between then two lineages could have occurred. However, the results presented here suggest that earlier introgression was not major, as the intact chromosomes of the southern breeds are predominantly taurine, while those in the horn are predominantly indicine in character. A similar situation is observed on examination of Y-chromosome data. Ogaden and Arashie, the breeds located closest to the horn of Africa have a private *B. indicus* allele at a diagnostic bi-allelic Y-chromosome locus, whereas the *B. taurus* allele predominates in the other breeds (HANOTTE *et al.* 2000).

The borders of the Sahel are not defined in either political or geographic terms, but by mean annual rainfall. This 300km wide region, which stretches across 5000 km of northern Africa along the southern fringes of the Sahara desert, exhibits a relative homogeneity of climate and agriculture, which defies national boundaries (**Figure 4.29**). However, there is great linguistic diversity, a result of the mixing of many ethnic groups including light-skinned Berber and Arabic people and dark-skinned people from the Sahara (ISNAR 1995).

Figure 4.29 *Map of Africa showing the sub-Saharan Sahel region. The northern and southern Sahel are marked in light and dark grey respectively.*

Archaeological evidence in West Africa was so sparse that the 3000 years before historical records have been called "the silent millennia" (McIntosh and McIntosh 1979). However, studies have revealed that trade routes through the Sahara have been used since ancient times; that the Sahara desert has bridged, rather than divided societies from both sides of the African continent. While the trade across the continent moved vast quantities of goods, presumably including domestic cattle, its impact was significant in other ways. People and ideas were exchanged and later Islam spread rapidly to the region (ISNAR 1995). MacDonald has hypothesised that there were "Mobile Elites" in the area from around 4000 BP. These were mobile herders who accumulated wealth through pastoralism (MacDonald 1998). While archaeological evidence will be skewed in favour of imports, the discovery and dating of glass beads and semi-precious stones from Asia shows that transcontinental movement was taking place 2000 years ago (Connah 2001).

Most authors agree that trade was widespread in West Africa even before importing from Arabia began. As Connah (2001) writes, "each environment possessed some resources but lacked others. Thus salt was available in the desert and along the coast but was relatively difficult to obtain in the savanna, where for cereal agriculturalists it was a physiological necessity. Thus the forest was deficient in meat, but the savanna supported very large numbers of domestic animals, particularly cattle." Therefore, the complexity of the West African environment, as a whole, provided conditions conducive to the development of a complex network of regional trade. Salt was mined extensively in Idjil (Mauritania), Taodeni (Mali), Fachi, and Bilma (Niger) (ISNAR 1995), but there is only archaeological evidence for the essential salt trade from the 1st millennium AD. Despite this, these trade networks are likely to have existed in the area as long as food production, which was well established by 3000 BP (CONNAH 2001). Therefore, Arab trade was able to link into these extant sub-Saharan trading routes and it was this that enabled its rapid establishment and expansion (MCINTOSH and MCINTOSH 1980). So extensive was the trade along the Sahel that the settlements that sprang up have been called "ports" on the edge of a "sea of sand" (LEVTZION 1973).

This fluid movement of people, animals and other goods has continued since ancient times. As the Sahara became gradually drier (c. 2500 BP) people moved to the south and many settled close to the Senegal and Niger rivers and around Lake Chad. Trade along the Sahalian corridor in agricultural produce and other goods remained the main means of contact for the different ethnic groups that were established. Cultivated plants such as pearl millet, cotton, indigo and rice were produced, and groups, such as the Peul, kept zebu cattle, sheep and goats (ISNAR 1995). The "Pulse Model" describes the tendency of the

Sahara to recede and expand in waves, causing the inhabitants of its extremities to uproot and move on a regular basis. Hence, the movement of people and livelihoods has remained a feature of this part of Africa (McIntosh 1993).

This picture of an ancient and dynamic intra-continental trading network contrasts somewhat with the apparent situation in eastern and southern Africa, where the coastal trade with Asia dominates (Connah 2001). Nevertheless, internal trading networks were almost certainly present as well. Initially described as the Swahili Corridor (Horton 1987) more recent authors have described a "Swahili Trading Wheel" (Sinclair 1995), characterised by a moving hub which connected the East African islands and coastal strip to the Indian Ocean trade network. However, the movement of cattle along these trading routes was certainly more recent (Marshall and Hildebrand 2002; Smith 2000), and possibly less dynamic than their movement along the Sahal corridor. In fact, several important differences can explain the contrasting patterns of the spread of agriculture in the north and south of Africa.

The northern Sahelian grasslands were the home to *B. primigenius*, therefore cattle are well adapted to this region and this enabled a rapid spread of pastoralism. Early herders would have had extensive links with hunter-gatherers, and also exploited local plants. The erratic rainfall of the Saharan region would have meant that mobility was essential to survival. The extensive trade routes and presence of water in the Sudanese region would also have contributed to the movements of pastoralists along the Sahel corridor, between ephemeral seasonal settlements (Marshall and Hildebrand 2002).

By contrast the early Iron Age groups in southern African lived in fairly permanent settlements. Despite the presence of arid areas, they did not have to contend with the changing rainfall patterns of the north. Additionally, livestock diseases, rainfall and desert areas kept farming confined to the eastern part of southern Africa (GIFFORD-GONZALEZ 2000). Movement of cattle was also restricted by their poorer adaptation to this novel habitat, which never accommodated any of their wild relatives (DIAMOND 1997). The earliest pastoralists of southern Africa were part of a fully developed Iron Age when they arrived in the region, and had little relationship with the local hunter-gatherers and did not domesticate local plants. As a result, cattle production was already integrated with crop production by the time it spread south and would have been more sedentary than northern pastoralism (MARSHALL and HILDEBRAND 2002).

The results presented here are concurrent with the two different pictures of pastoralism in the north and south of Africa. The high level of recombination within haplotypes in the north is indicative of old admixture, and may be the result of an old and dynamic culture of pastoralism through the Sahel that was profoundly affected by *B. indicus* introgression almost as soon as zebu entered the African continent. Hybridisation is likely to have taken place between both older and newer hybrid populations with flexible movement between groups, facilitated by well-established trade routes. By contrast, southern pastoralism is a more recent development and probably less mobile. A similar picture emerges of two contrasting regions within the Near East (**Figure 4.23**), where earlier introgression into the Iraqi breed has produced more detectable recombination than is seen in Anatolian Black.

4.4.4 APPLICATIONS AND FUTURE WORK

Understanding how *B. indicus* introgression into African taurine cattle has affected trypanotolerance in some populations is critical if this important genetic resource is to be preserved. Admixture studies can quantify this effect and may also be useful in identifying the genes responsible for this disease resistance. Additionally, it is important to estimate background levels of LD in different livestock populations before any mapping studies are initiated, to avoid false positives associations. Such studies should provide estimates of the spacing of markers that will be required to convince that an association between a gene of interest and a neutral marker is genuine and to account for genome heterogeneity (COLLINS *et al.* 1999; FREIMER *et al.* 1997; PETERSON *et al.* 1995; WILSON and GOLDSTEIN 2000).

An inherent difficulty in the quest to map genes for complex traits in humans using admixed populations is the lack of diversity in *Homo sapiens*. Successful admixture mapping requires correct assignment of ancestry to an allele, a challenging task when there are few population-specific alleles. Lonjou et al. (1999) have even hypothesised that there are no significant differences in LD levels between human populations, but this is probably not the case (KRUGLYAK 1999a; LONJOU *et al.* 1999). Ancestry informative markers are relatively rare in human populations, however many alleles distinguish between the between taurine and indicine bovine lineages. There is a range of existing hybrid populations and experimental crosses can even be carried out. There are also traits, such as trypanotolerance, which occur at varying levels according to levels of admixture. These populations are ideal candidates for mapping using admixture and linkage disequilibrium. Therefore it may prove to be more straightforward to map genes in admixed cattle populations where the proportion of admixture in an individual can be easily estimated and there are numerous population-specific alleles. A genome scan of one marker every 5-

10 cM may even be sufficient to detect significant associations with careful population selection (KAPLAN *et al.* 1998; LAUTENBERGER *et al.* 2000; MCKEIGUE 1997; WILSON and GOLDSTEIN 2000).

Significant associations between linked markers may be useful in neutral marker mapping, as a very significant association may correlate with close physical linkage (WU and ZENG 2001). As LD decays with time, it may be possible to use admixture and LD studies on to date when a hybridisation event occurred (HUTTLEY *et al.* 1999; OTT 2000). The significance level of an association between two mapped loci a certain distance apart may be used as a measure of when the hybridisation event took place (MCKEIGUE 1998; TENESA *et al.* 2003).

CHAPTER 5

GENETIC ANALYSIS USING MULTIPLE MICROSATELLITE DATASETS

5.1 INTRODUCTION

Following the assessment of diversity and admixture in regional cattle populations detailed in previous chapters, a reasonable aim was to try and obtain a global picture of bovine admixture and diversity using three extensive microsatellite datasets consisting of samples from different but overlapping geographic areas. Two methods were used. Firstly, a fully Bayesian analysis of admixture was implemented for each dataset, inferring different numbers of putative parental populations. This analysis aims to address the question of structure within and between *a priori* defined groups and investigate whether data from different microsatellite markers will reveal the same underlying pattern of admixture. Secondly a regression-based method was employed to combine optimum admixture and diversity estimates into one dataset covering the entire geographic area contained within the three original groups.

5.1.1 DATA

Domestic cattle represent an interesting subject to demonstrate data combining methods, as there are three large microsatellite datasets available which contain some common breed and markers. Furthermore, the history of domestic cattle has been quite thoroughly investigated; so prior knowledge enables us to test the accuracy of our method. Previous analyses show that modern domestic cattle are most likely derived from two or three wild progenitor groups. Domestication is thought to have taken place in the Near East, the Indo-Pakistan region and probably north Africa. Previous work has demonstrated that the genetic signature of these different wild regional variants of aurochs remains intact in modern populations (**Chapter 1**).

The three datasets that are included in the current study cover a wide geographical scope. The markers in the earliest dataset (Dataset 2) were developed directly from the public domain sequence database, and hence some are closely linked to genes. These twenty microsatellite markers are among the earliest bovine microsatellites, and two are located within the bovine major histocompatibility complex. This set was used in some of the first published population genetics studies that employed microsatellites, and is mainly focussed on samples from West Africa (MACHUGH *et al.* 1994; MACHUGH *et al.* 1997). Control samples from India, Europe and East Africa are also included.

The largest dataset (Dataset 1), in terms of sample number, is concentrated on the African continent. Fifteen markers have been genotyped in a range of breeds that span most of the continent where cattle are found. Reference breeds from India and Europe were also included within this set. The final dataset (Dataset 3) has the broadest geographical range and is comprised of twenty markers that have been recommended for collaborative study (see **Chapter 2**). This dataset contains the only cattle samples from the Near East, many breeds from both Europe and India, but a sole African breed.

The existing global cattle population is the result of migration, breeding and admixture that has taken place since domestication in three main centres (**Chapter 1**). Broadly speaking, *B. indicus* cattle from Asia have migrated into the Near East and large areas of Eastern and Southern Africa while taurine cattle have moved from the Near East and the north of Africa to populate the Eurasian landmass and Africa respectively. None of the previously mentioned datasets include samples from all putative centres of domestication and important areas of secondary introgression. Collectively, however, most significant regions are included.

5.1.2 BAYESIAN ANALYSIS OF ADMIXTURE

A drawback of a traditional population genetics sampling and genotyping approach is that the global population under study must be divided into sub-populations or demes. In this case, sampling location, breed and phenotype have been used as dividing attributes. Obviously, subdivision has an important impact on subsequent examination of populations sub-structuring and may or may not reflect underlying genetic relationships (ROSENBERG *et al.* 2002). If a deme is split into two different sample sub-populations no structuring will be evident between those groups (BALLOUX and LUGON-MOULIN 2002). However, if a sample is not homogeneous, and in reality consists of several demes, it may be estimated that that the sample contains admixed individuals or that there is more within-sample variation than is actually the case.

Recently developed Bayesian methods allow researchers to carry out "blind" admixture studies in hybrid and introgressed populations. In essence, it is not necessary to define which of the populations in the analysis might best represent the original contributors. This method represents a step forward from traditional admixture analysis, where there has always been an inherent bias, as parental populations were selected based on previous scientific results or historical information (BIERNE *et al.* 2003; ESTOUP *et al.* 1999). If this method is implemented on different microsatellite data sets and finds that the same populations cluster on the basis of distinct allele distributions, this is strong evidence that the method detects a genuine underlying admixture pattern. This analysis is performed using the STRUCTURE program (see **Chapter 2**), which employs a probabilistic assignment of admixture proportions to individuals (PRITCHARD *et al.* 2000a).

Rosenberg and colleagues (2002) implemented this method to examine partitioning between human populations groups and found that, in general, predefined populations

usually corresponded to genuine genetically distinguishable groups. Clusters were modelled to identify sub-groups with distinctive patterns of allele frequencies. The authors found that individuals from the same predefined population nearly always shared membership coefficients in inferred clusters. Geographic groupings seemed to correspond to inferred groupings when up to a certain number of clusters were inferred. After this, when a greater number of populations was inferred, individual human populations (which may be defined by language and/or location), were identified as separate categories. Therefore, major groupings may be identified when a lower number of clusters are implied, whereas inferring a large number of distinct populations may reveal smaller discrete genetic units. Genetic structuring reflects the number of alleles that have been exchanged between populations and hence has profound effects on the composition of individuals within those populations. Estimates of population differentiation are essential to understanding relationships among populations and may be helpful in deciphering past contacts or recent divergence (BALLOUX and LUGON-MOULIN 2002).

5.1.3 COMBINING DATA

One of the obstacles to producing comprehensive maps of diversity and admixture components has been the non-uniformity of different microsatellite datasets. When different markers, with differing levels of polymorphism, are genotyped in different populations the potential to extract useful information between and among datasets is quite limited. In any sample with any given set of markers one obtains an estimate of the true value (of, say, the admixture proportion) of the population. As the sample size increases and the number of markers increases, the estimate of the admixture level should tend toward the true value of admixture for that population.

It is assumed that estimates of admixture will asymptote to the true admixture proportion as the number of markers increases (MOAZAMI-GOUDARZI and LALOE 2002). If one had data on true admixture proportions, from two different but partially overlapping data sets, one could estimate directly the relationship between the two sets of admixture proportions and convert from one data set to the other, allowing direct combining of the two sets of admixture data into a single data set. The admixture values for the populations in common would be the same for each marker set, and if regression analyses were performed with these actual values, it would be expected that the resulting line would have a slope of unity. In practice one has "estimates" of admixture from two data sets and the regression of one set of estimates on the other does not provide an unbiased conversion from one data set to the other.

As more markers are added to any set of data, estimates made are getting closer to the true value of any variable (χ), whether it be admixture or diversity measures. If it was possible to predict the actual value of these variables with different datasets and different populations, different datasets could be combined directly. This is not the case as all

current datasets are fairly small numbers of markers. Here, there were significant correlations between different datasets using point estimates of several variables (such as admixture measures) measured in populations held in common. Common populations can therefore be used in regression analysis to elucidate the relationship between the (imperfect) estimates made using the two datasets. Linear transformation can be used to transform every population from each dataset to a best-fit line and these transformed values become our estimates of the variable in question.

Here, data from the three groups of microsatellites are combined for two diversity measures: mean number of alleles ($\bar{A}$) and unbiased heterozygosity ($\hat{H}$), and for admixture levels when three parental populations are inferred by the Bayesian method. Combination of admixture and diversity measures from multiple datasets is feasible, providing there are populations in common between them. Ideally these common populations or breeds should span a range of admixture proportions and diversity levels. Obviously, the greater the number of breeds held in common between the datasets, the more accurate the final diversity and admixture analyses.

Table 5.1 *(part 1) List of breed population sizes (N) and average predicted diversity ($\hat{H}$ and $\bar{A}$) and admixture variables (q_e, q_i and q_a). Datasets 1, 2 and 3 are denominated DS1, DS2 and DS3 respectively.*

Population	N (DS1)	N (DS2)	N (DS3)	Ave $\hat{H}$	Ave $\bar{A}$	Ave q_e	Ave q_i	Ave q_a
Abigar	35			0.658	4.680	0.030	0.725	0.256
Afar	35			0.600	4.352	0.021	0.909	0.087
Afrikaner	35			0.547	3.091	0.026	0.136	0.827
Alentejana			30	0.672	5.031	0.884	0.017	0.136
Anatolian Black	34		41	0.735	5.881	0.459	0.413	0.129
Angoni	35			0.655	4.610	0.049	0.696	0.263
Angus		33		0.447	2.409	0.996	0.013	0.031
Ankole	35			0.574	3.842	0.045	0.750	0.216
Arado	35			0.638	4.465	0.027	0.947	0.044
Arashie Beja	34			0.607	4.342	0.027	0.938	0.053
Arouquesa			34	0.686	5.091	0.917	0.019	0.101
Arsi	35			0.658	4.729	0.021	0.922	0.075
Azawak		29		0.568	3.143	0.054	0.859	0.074
Baladi/Egyption	40		32	0.720	5.693	0.429	0.328	0.246
Bale	32			0.666	4.531	0.027	0.952	0.039
Baoule 1	35	56		0.491	3.035	0.133	0.019	0.854
Baoule 2		46		0.512	3.200	0.061	0.077	0.883
Baoule 3		50		0.529	3.088	0.036	0.083	0.901
Barotse	35			0.584	3.785	0.021	0.061	0.904
Blonde	35			0.531	3.393	0.856	0.016	0.094
Boran Ethiopia	35			0.657	4.603	0.014	0.969	0.036
Borgou		60		0.614	3.621	0.068	0.406	0.533
Butana		24		0.582	2.978	0.026	0.879	0.081
Butana Foya	33			0.619	4.335	0.091	0.847	0.075
Charolais	33	36	36	0.608	3.928	0.911	0.012	0.078
Damascus			41	0.752	5.800	0.306	0.354	0.353
Danakil	35			0.635	4.716	0.018	0.959	0.042
Desi			26	0.751	5.828	0.034	0.937	0.013
East Anatolian Red			44	0.799	6.991	0.494	0.436	0.082
Evolenard			15	0.599	4.030	0.983	0.010	0.046
Fogera	35			0.641	4.817	0.023	0.915	0.079
Friesian	35	40		0.551	3.502	0.966	0.013	0.021
Fulani		24		0.621	3.268	0.023	0.889	0.074
Gobra	35	59		0.634	4.282	0.075	0.660	0.268
Hariana		10	10	0.637	3.933	0.015	0.956	0.011
Hereford		34	34	0.521	3.186	1.011	0.011	0.017
Highland			37	0.523	3.887	0.989	0.017	0.032
Highland Zebu	33			0.661	4.942	0.028	0.944	0.047
Horro	29			0.670	4.935	0.018	0.883	0.116
Hungarian Grey			44	0.636	4.515	0.911	0.033	0.094
Iraqi			42	0.807	7.211	0.190	0.773	0.030
Iringared	35			0.635	4.410	0.021	0.919	0.078
Jersey	35	34	44	0.537	3.397	0.919	0.024	0.058
Kapsiki	35	60		0.583	3.790	0.093	0.241	0.671
Kavango	35			0.591	3.807	0.035	0.122	0.831
Kavirondo	35			0.654	4.459	0.033	0.886	0.098
Kenana		38		0.603	3.219	0.017	0.942	0.024
Kenyan Boran	35			0.690	4.865	0.080	0.868	0.066
Kerry		40		0.451	2.222	0.975	0.014	0.050
Kigezi	35			0.633	4.196	0.029	0.465	0.506
Kilimanjaro	35			0.641	4.644	0.023	0.926	0.069
Koakoland	35			0.617	4.130	0.029	0.094	0.863

Table 5.1 (part 2).

Population	N (DS1)	N (DS2)	N (DS3)	Ave $\hat{H}$	Ave $\overline{A}$	Ave q_e	Ave q_i	Ave q_a
Kurdi			37	0.812	7.150	0.351	0.573	0.080
Kuri	35	52		0.626	4.185	0.061	0.416	0.527
Laguna		60		0.444	2.686	0.039	0.055	0.928
Landim	35			0.581	4.047	0.030	0.105	0.853
Malawi Zebu	35			0.633	4.566	0.032	0.897	0.088
Marmarena			40	0.682	5.035	0.815	0.153	0.062
Mashona	35			0.632	4.543	0.085	0.073	0.827
Maure	35	55		0.650	4.331	0.057	0.783	0.163
Mbororo	35	47		0.618	3.905	0.050	0.795	0.158
Mertolenga			37	0.670	5.018	0.951	0.022	0.065
Modicana			37	0.655	5.186	0.835	0.153	0.042
Muturu		28		0.466	2.621	0.031	0.090	0.899
Namchi	35	54		0.548	3.661	0.047	0.108	0.851
N'Dama (Gambia)		58		0.566	3.405	0.041	0.095	0.884
N'Dama (Guinea Bissau)		54		0.514	2.917	0.092	0.021	0.911
N'Dama (Guinea)	35	63	44	0.504	3.162	0.052	0.011	0.944
N'Dama (Mali)		44		0.556	3.197	0.052	0.081	0.888
N'Dama (Nigeria)		19		0.538	3.128	0.048	0.092	0.881
N'Dama (Senegal)		48		0.531	3.226	0.052	0.050	0.921
Nelore	35		27	0.633	4.529	0.025	0.968	0.010
Nguni	35			0.605	4.243	0.024	0.057	0.905
Nkone	35			0.622	4.335	0.172	0.201	0.614
Nuba	35			0.624	4.883	0.035	0.840	0.139
Ogaden	35			0.652	4.412	0.023	0.962	0.034
Old Brown			50	0.657	4.769	0.993	0.013	0.034
Ongole	28		32	0.630	4.295	0.024	0.965	0.013
Orma Boran	30			0.669	4.325	0.062	0.901	0.053
Pedi	35			0.577	3.511	0.059	0.045	0.879
Peul Fulani			79	0.598	3.634	0.046	0.803	0.141
Pinsguar			30	0.718	5.618	0.943	0.028	0.067
Pushtetaller			31	0.701	5.035	0.927	0.046	0.064
Rayazebu	35			0.639	4.454	0.019	0.949	0.051
Red Sindhi			38	0.713	5.540	0.059	0.909	0.017
Retinta	32			0.568	3.465	0.920	0.021	0.011
Romagnola			40	0.617	4.645	0.855	0.097	0.081
Sahiwal	35	13	10	0.621	4.053	0.025	0.967	0.010
Saler			38	0.552	4.296	0.990	0.017	0.032
Sheko	35			0.685	4.913	0.044	0.805	0.164
Simmenthal		36	50	0.521	3.469	1.005	0.009	0.025
Sokoto Gudali	35	47		0.622	3.911	0.044	0.777	0.182
Somba	35	60		0.528	3.389	0.086	0.023	0.897
South Anatolian Red			43	0.798	6.813	0.510	0.362	0.145
Sykia			21	0.704	4.966	0.851	0.033	0.152
Tharparker		10	35	0.595	3.441	0.016	0.956	0.011
Tonga			35	0.599	3.933	0.025	0.062	0.899
Tuli			34	0.620	3.795	0.107	0.140	0.739
Turkish Grey			43	0.778	6.796	0.790	0.152	0.087
Vosges			27	0.696	4.873	0.993	0.016	0.030
Watusi			35	0.615	3.687	0.033	0.525	0.444
White Fulani			30	0.600	3.290	0.023	0.914	0.047
Zebu Malagasy			35	0.535	3.460	0.014	0.967	0.039

5.2.2 CALCULATION OF ADMIXTURE FOR MULTIPLE VALUES OF k.

The Bayesian admixture analysis was implemented using the program STRUCTURE *v.* 2.0 (PRITCHARD *et al.* 2000a). Each dataset was analysed separately using the admixture model, a burn-in of 200,000 iterations and 100,000 repetitions of the Markov chain. The number of clusters that are to be inferred is denoted k. Each dataset was analysed twice inferring a range of values of k. The range of values that was used for each dataset is shown in **Figure 5.1**. In highly structured data, the most divergent groups tend to divide into distinct clusters first (PRITCHARD *et al.* 2000a). Less divergent groups will separate as the value of k increases, although sample sizes and within-group diversity levels also affect the order in which clusters split (FALUSH *et al.* 2003; ROSENBERG *et al.* 2001a). The authors of STRUCTURE recommend running the program for many values of k and selecting the model that has the maximum posterior probability (PRITCHARD *et al.* 2000a). However, differences between the likelihood estimations for different models may be quite trivial, and different models may also reveal a range of levels of population structure (ROSENBERG *et al.* 2002). Graphical displays of the results were produced using DISTRUCT, which is available at http://www-hto.usc.edu/~noahr/distruct.html (ROSENBERG *et al.* 2002). Plots of both individual and breed admixture levels were produced.

5.2.3 COMBINATION OF DIVERSITY AND ADMIXTURE MEASURES

5.2.3.a Calculation of variables for combination

$\overline{A}$ and $\hat{H}$ values were calculated for datasets 1, 2 and 3 using **Equations 2.1 and 2.2** respectively. Admixture coefficients were calculated for datasets 1 and 4 using the STRUCTURE program inferring three parental populations. Datasets 2 and 3 were pooled for this part of the analysis as the majority of individuals in dataset 2 are admixed (STRUCTURE *v.* 2.0, users guide). Dataset 3 contains only one indigenous African breed, the individuals of which represent a putative African parental population. An admixture

model implying three parental populations will not be accurately inferred by this dataset when one of the parental populations is characterised by African taurine cattle (see Section 5.3). To circumvent this problem, datasets 2 and 3 were pooled for the seven loci held in common between them. This combined dataset was denoted dataset 4 and was subsequently combined with admixture estimates for dataset 1.

Three admixture coefficients were calculated for each individual in the two relevant datasets (1 & 4). STRUCTURE was set to infer three parental populations and run for a burn-in period of 100,000 steps and then for 1,000,000 iterations of the Monte Carlo Markov Chain. Each coefficient in the output represents the probable amount ancestry of an individual derived from each of three inferred parental populations. These values were averaged over breeds to get three $\hat{q}$ values each. These were denoted q_e, q_a and q_i for the populations best characterised by European/Near Eastern taurine, African taurine and Indian *B. indicus* respectively. The diversity and admixture results for the different datasets were combined using the method described in the following section.

5.2.3.b Combination of variables

5.2.3.b.i The Method

The method can be applied to two datasets which have been genotyped for different sets of microsatellites markers and which have some populations (in this case, breeds) in common. A variable, such as $\overline{A}$ or an admixture coefficient, is calculated for each population in each data set. A regression analysis is carried out using the value in question and populations held in common between the two datasets. It is essential to obtain highly significant *p*-values and high r^2 (adjusted) values for the regressions analyses. These give confidence that there is a genuine relationship between the variable estimates for the two datasets.

1. Given a variable in two datasets denoted χ in dataset I and γ in dataset II, two regressions are carried out using populations that are found in both of the datasets. Each regression is carried out with each dataset as dependent and independent variable.

$$\beta(\chi,\gamma) = Cov(\chi,\gamma)/Var(\gamma)$$

$$\beta(\gamma,\chi) = Cov(\chi,\gamma)/Var(\chi)$$

2. The equations of the two lines from these regressions will be:

$$\chi = \gamma b + a$$

$$\gamma = \chi c + d$$

where b and c are the slopes of the lines and a and d are constants. The reciprocal of each line can be calculated as follows:

$$\gamma = (\chi - a)/b$$

$$\chi = (\gamma - d)/c$$

In a case where there are no errors associated with the regression the reciprocal lines will be identical as errors are minimised on the vertical using the least squares methods. In the case of most data the original and reciprocal lines are not identical so we take the average of the two lines.

Average of $\chi = \gamma b + a$ and $\chi = (\gamma - d)/c$ is $\chi = \gamma h + i$

Average of $\gamma = \chi c + d$ and $\gamma = (\chi - a)/b$ is $\gamma = \chi f + g$

3. Two predicted datasets (denominated $pred\chi$ and $pred\gamma$) are created by transformation using these average lines.

$pred\chi = \gamma h + i$ when there is a γ value in dataset II and $pred\chi = \chi + i$ when there is no γ value in dataset II.

$pred\gamma = \chi f + g$ when there is a χ value in dataset I and $pred\gamma = \gamma + g$ when there is no γ value in dataset I.

This will produce two new datasets (predicted χ and predicted γ) that have values for all the populations in both original datasets. A second regression analysis is carried out using the predicted values for all breeds.

$$\beta(pred\chi, pred\gamma) = Cov(pred\chi, pred\gamma) / Var(pred\gamma)$$

$$\beta(pred\gamma, pred\chi) = Cov(pred\chi, pred\chi) / Var(pred\gamma)$$

New linear transformation equations are calculated by repeating step 2. The predicted datasets are then transformed a second time to minimise errors in the regression and to get closer to the average best-fit line for the data (new predicted values). The two new predicted values for each population are then averaged to get an average predicated value that can be visualised on a contour map.

5.2.3.b.ii Three datasets

To combine three datasets, pairwise regressions were carried out between all possible combinations. The two datasets with the highest correlation (as determined by the regression parameters) were combined using the method outlined above. The new predicted dataset produced was then combined with the third dataset. This method was used for the $\overline{A}$ and $\hat{H}$ transformations. The highest original correlation in both cases was between datasets 1 and 2. The higher number of populations in common between datasets 1 and 2 (n = 14) may explain this. Diversity measures (especially $\hat{H}$) may not span a large range between different populations and unless there are enough representatives from the full range of values in the common breeds it may not be appropriate or possible to determine the relationship between the two datasets using regression analysis.

5.2.3.c Visualisation of transformed combined data.

Longitude and latitude values for each breed were chosen using breed origin information from Felius (1995). The ARCVIEW GIS program (*v.* 3.2, Environmental Systems Research Institute, ESRI, Redlands, CA) was used to plot the average predicted values for unbiased heterozygosity, mean number of alleles and the admixture coefficient values for each breed

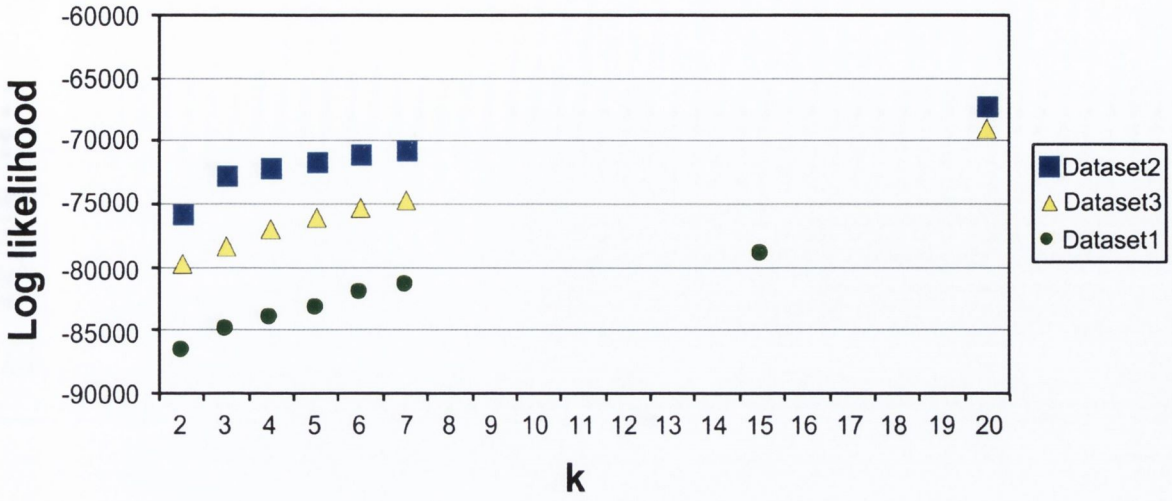
at the point of its geographical origin. A projection of 1:83,000,000 and 20 equal interval contours were used to visualise the data. Maps were produced using the Adobe ILLUSTRATOR package.

5.3 RESULTS

5.3.1 POPULATION STRUCTURE WITHIN DATASETS

Each of the three datasets were analysed independently for admixture levels using STRUCTURE and a range of different values of k. The values of k that were used are shown in **Figure 5.1**. While adding degrees of freedom to a model will tend to increase likelihood, the probabilities begin to asymptote at a k value, which should explain the major structure in the data. For the current datasets 1, 2 and 3, these values seem to be 3, 3 and 4 respectively. No prior phenotypic or location information was used in the clustering analysis. Results are shown in **Figures 5.2 to 5.7**. Each figure shows one of the datasets and each rectangular plot represents the divisions produced by a given value of k. The upper plot shows a k value of two, and each successive plot below shows the divisions in the data when k is increased by one, to a maximum value of seven in the lowermost plot. **Figures 5.2, 5.4** and **5.6** show individual levels of admixture. In these ragged plots, each vertical line represents one individual and the black vertical lines show breed divisions. Breed codes are shown beneath each plot (**Appendix B**) while phenotype and geographic information is shown above. T, Z, S and H stand for *B. taurus*, *B. indicus*, Sanga and taurine/indicine hybrid respectively (**Chapter 1**). **Figures 5.3, 5.5** and **5.7** show the average breed admixture coefficients, where the vertical black dividing lines within the rectangular plots separate breeds.

Figure 5.1 *Plot of log likelihood values of different k values for each of the three current datasets.*



5.3.1.a Dataset 1

The clustering results for Dataset 1 are shown in **Figures 5.2** and **5.3** for individuals and breeds respectively. A summary of the clustering results is shown in **Table 5.2a**. When dataset 1 is divided by a k value of 2 there is a clear split between the ancestral bovine lineages, *B. taurus* (blue) and *B. indicus* (yellow). The next division separates the taurine constituent into two and seems to highlight a southern African component to the data (green). In the four-way division, it is African taurine that are shown in green, European/Near Eastern taurine are represented by blue, and two indicine components are evident. Yellow still shows the Indian *B. indicus* group, while red corresponds to an African indicine element. k=5 splits the African indicine component into two clusters, red and purple.

The next graph shows that European/Near Eastern taurine are in blue, West African taurine breeds are now shown in turquoise, while the southern African taurine component is still separated and is coloured green. Indian zebu cattle breeds are shown in yellow, while the two components of African *B. indicus* are still highlighted in red and purple. The final division of the breeds (k=7) is the same as the previous divisions, but a fourth division

(coloured grey) is now seen within Africa, which primarily groups some indicine/sanga breeds from the lakes area of Uganda and Malawi.

Table 5.2a *Interpretation of clusters produced by Bayesian admixture analysis of Dataset 1. Data is presented in* ***Figures 5.2 and 5.3****. Abbreviations used are SA=Southern African, WA=West African, Eur=Europe.*

k/colour	***Yellow***	***Blue***	***Green***	***Red***	***Purple***	***Turquoise***	***Grey***
2	Indian *B. indicus*	*B. taurus*					
3	Indian *B. indicus*	*B. taurus*	SA *B. taurus*				
4	Indian *B. indicus*	Eur *B. taurus*	African *B. taurus*	African *B. indicus*			
5	Indian *B. indicus*	Eur *B. taurus*	African *B. taurus*	African *B. indicus*	African *B. indicus*		
6	Indian *B. indicus*	Eur *B. taurus*	SA *B. taurus*	African *B. indicus*	African *B. indicus*	WA *B. taurus*	
7	Indian *B. indicus*	Eur *B. taurus*	SA *B. taurus*	African *B. indicus*	African *B. indicus*	WA *B. taurus*	African lakes *B. indicus*

Figure 5.2 *DISTRUCT plot showing individual admixture levels from STRUCTURE analysis of Dataset 1 with the number of inferred populations set from two to seven. Each vertical line represents one individual.*

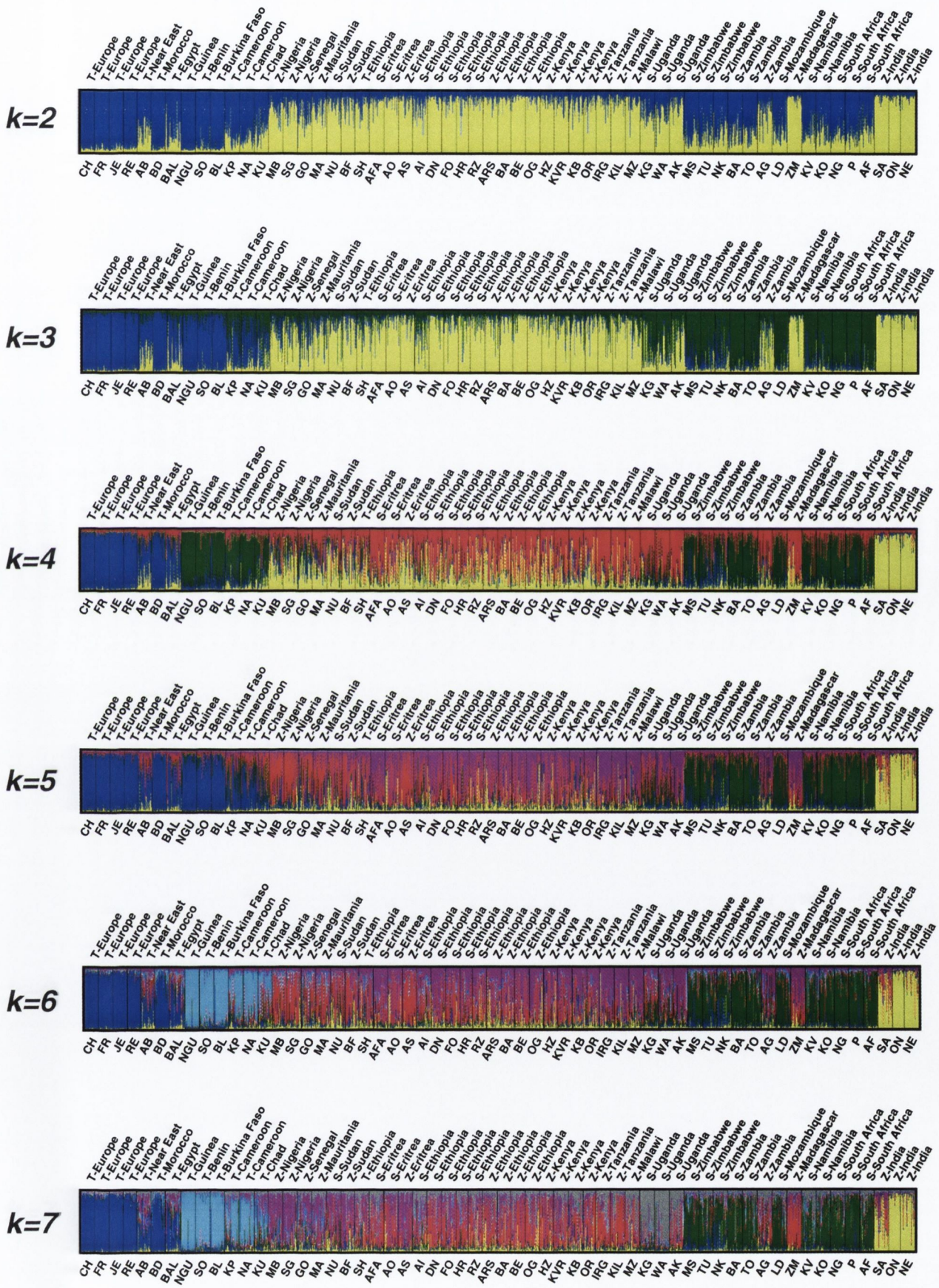
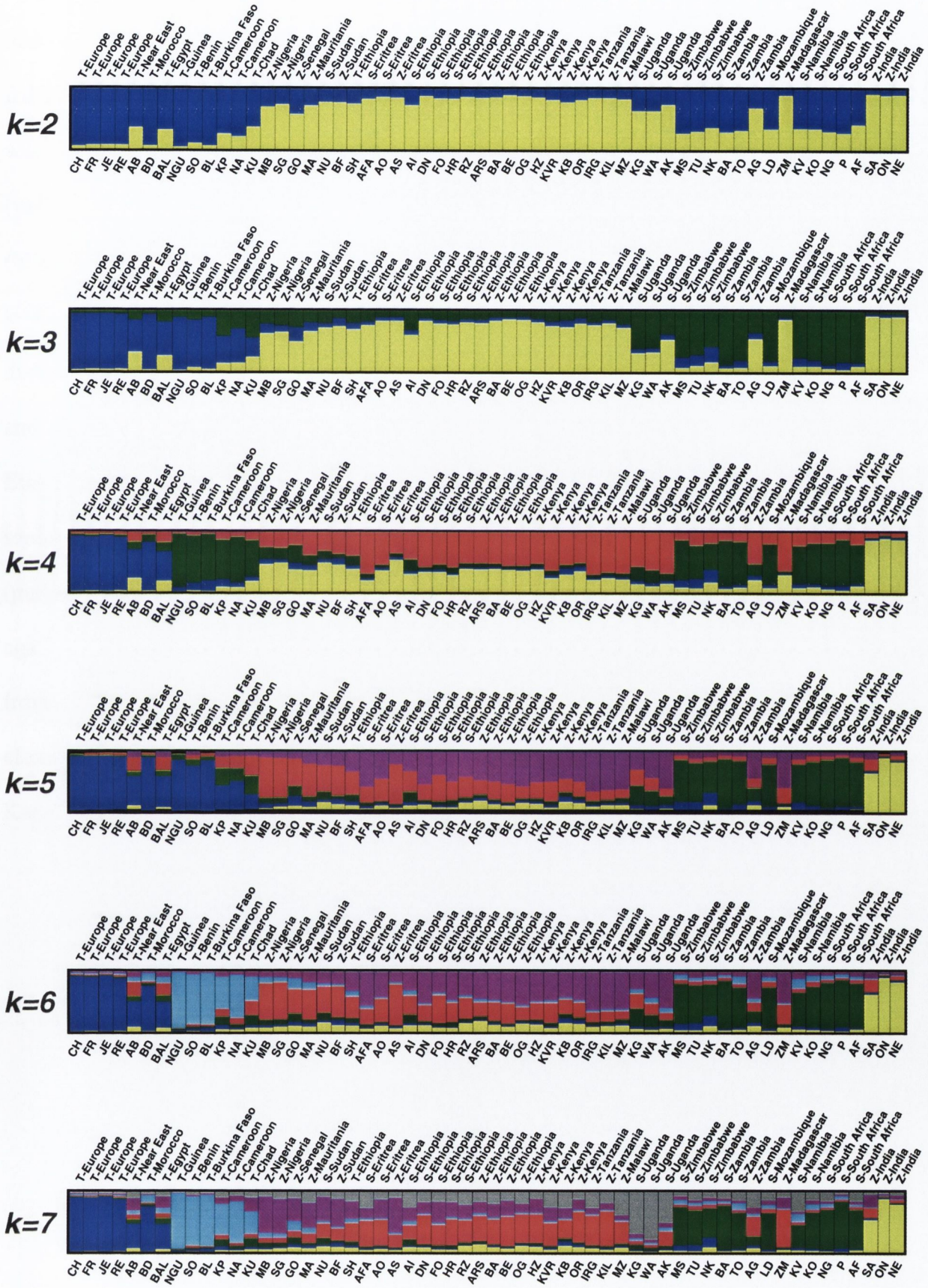


Figure 5.3 *DISTRUCT plot showing breed admixture levels from STRUCTURE analysis of Dataset 1 with the number of inferred populations set from two to seven.*



5.3.1.b Dataset 2

The clustering results for Dataset 2 are shown in **Figures 5.4** and **5.5** for individuals and breeds respectively. The clustering results are summarised in **Table 5.2b**. The three Indian breeds that are included in this dataset are Hariana, Sahiwal and Tharparker. As was the situation for dataset 1, the first division splits *B. taurus* (blue) and *B. indicus* (yellow). When k=3, the taurine element of the data is sub-divided into European taurine (blue) and African taurine (green). The African taurine cluster is further differentiated by k=4 and is shown by the green and red element where the N'Dama populations are distinguished from the others. k=5 introduces the African indicine component of the data and this is highlighted by turquoise. A six-way partition of the samples divides the European taurine individuals. The blue cluster still contains some of the European taurine samples, but some of the members of this regional group are now shown in a separate set (purple). In the k=7 graph, it can be seen that European *B. taurus* individuals are once again all grouped in the blue cluster, while the longhorn N'Dama (green) are differentiated from the West African shorthorn breeds (red). Purple is now colouring an African indicine element of the data, while a final division (grey), is found at high levels in Namchi, Kapsiki and Muturu, geographically proximal central African taurine breeds.

Figure 5.4 *DISTRUCT plot showing individual admixture levels from STRUCTURE analysis of Dataset 2 with the number of inferred populations set from two to seven.*

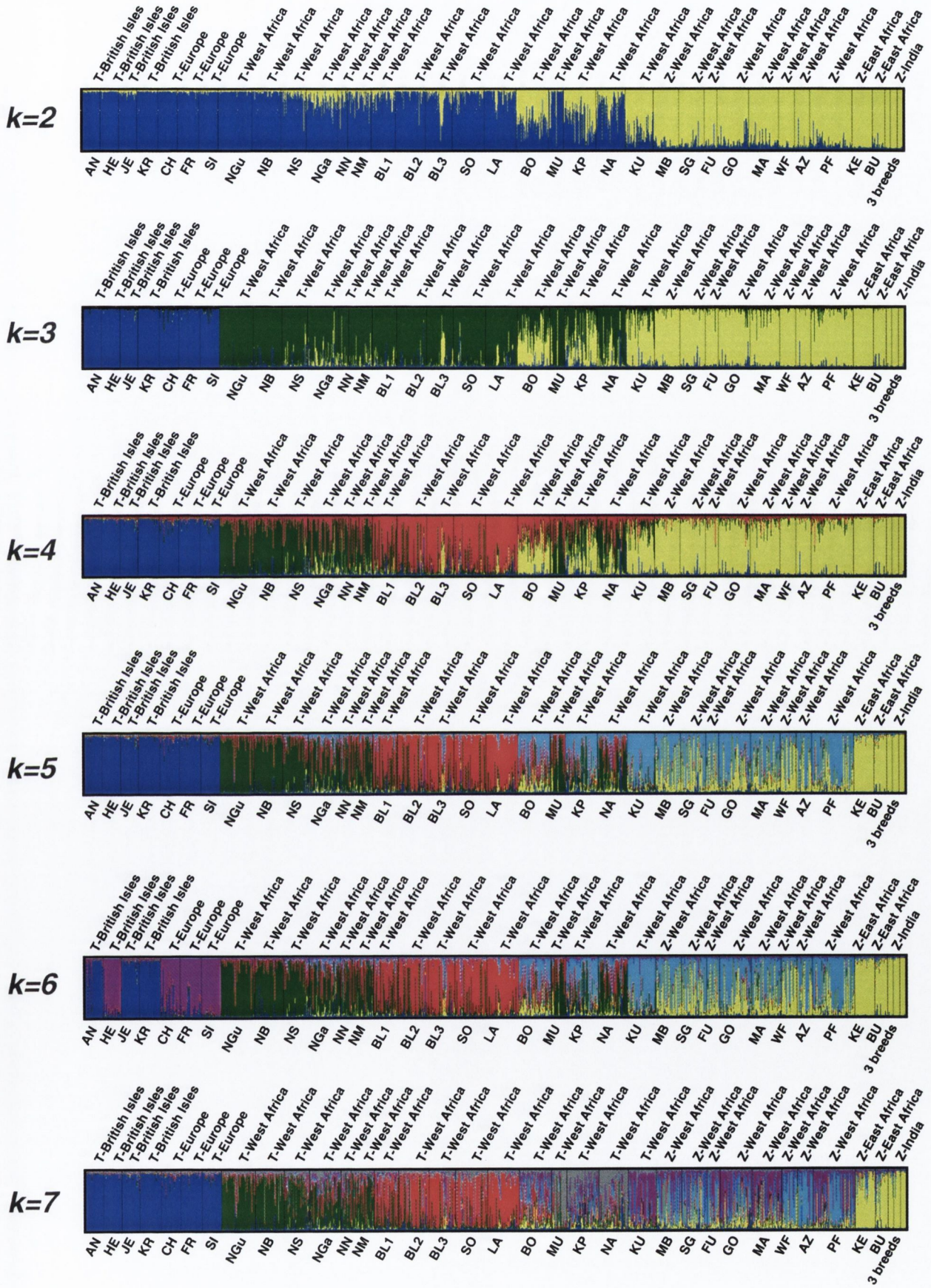


Figure 5.5 *DISTRUCT plot showing breed admixture levels from STRUCTURE analysis of Dataset 2 with the number of inferred populations set from two to seven.*

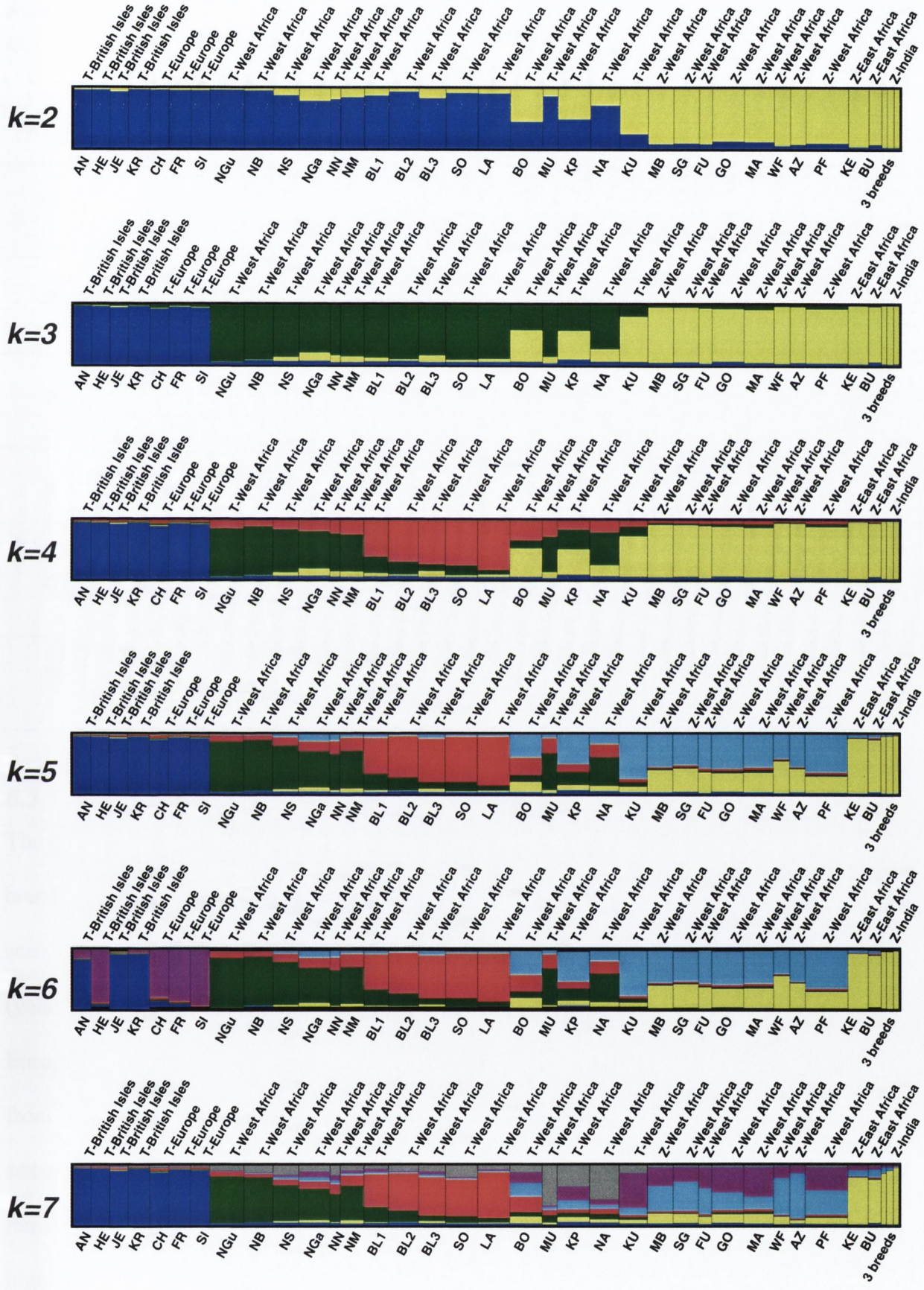


Table 5.2b *Interpretation of clusters produced by Bayesian admixture analysis of Dataset 2. Data is presented in* ***Figures 5.4 and 5.5****. Abbreviations used are UK=United Kingdom, Eur=Europe. Other abbreviations are cattle breed codes, which can be found in Appendix B.*

k/colour	*Yellow*	*Blue*	*Green*	*Red*	*Turquoise*	*Purple*	*Grey*
2	Indian *B. indicus*	*B. taurus*					
3	Indian *B. indicus*	Eur *B. taurus*	African *B. taurus*				
4	Indian *B. indicus*	Eur *B. taurus*	African *B. taurus*	African *B. taurus*			
5	Indian *B. indicus*	Eur *B. taurus*	African *B. taurus*	African *B. taurus*	African *B. indicus*		
6	Indian *B. indicus*	Some UK *B. taurus*	African *B. taurus*	African *B. taurus*	African *B. indicus*	Eur *B. taurus* and Hereford	
7	Indian *B. indicus*	Eur *B. taurus*	African *B. taurus longhorn*	African *B. taurus shorthorn*	African *B. indicus*	African *B. indicus*	WA *B. taurus* NA, KP, MU

5.3.1.c Dataset 3

The clustering results for Dataset 3 are shown in **Figures 5.6** and **5.7** for individuals and breeds respectively. A summary of the clustering results is shown in **Table 5.2c**. As was seen with the previous two datasets, the first division splits *B. taurus* (blue) and *B. indicus* (yellow). The UK, Mediterranean areas and Africa (green) are separated from other European breeds (blue) when k=3. This north central European group remains divided from other breeds throughout the analysis. k=4 subdivides the green cluster. Most of the same breeds continue to group and are coloured red, but Jersey (JE) and N'Dama (NGu) remain strongly classified together. When a further division of the data is carried out, blue highlights mid-European taurine breeds, green appears in African and some Near Eastern

taurine breeds, red is found predominantly in the Italian breeds Romagnola, Marmarena and Modicana and the Greek Sykia breed, but also in Highland and Hereford (UK). The element of the data shown by turquoise is seen strongly in the Portuguese breeds, Alentjana, Mertolengo and Arouquosa, but also in some other European breeds. At k=6, many of the divisions remain the same, or quite difficult to decipher. Green strongly delineates African *B. taurus*, but is also seen in the Near East. The same UK and Mediterranean breeds are shown by red. Other European breeds are shown by the turquoise set. Purple highlights Charolais and Jersey.

The Mediterranean breeds are divided between being strongly assigned to either the turquoise or purple grouping when k is set to 7. Italian, Greek (Sykia) and Hungarian breeds are shown in turquoise and cluster together while the Portuguese breeds group with the other mid-European breeds, including the French breeds, Charolais and Vosges and British Jersey breed (purple). Highland and Hereford are strongly assigned to a red grouping. Swiss, German and Dutch European breeds are coloured blue, and N'Dama is the only strong representative of African taurine (green). Grey shows a Near Eastern dimension of the data and Indian *B. indicus* breeds remain yellow.

k/colour	*Yellow*	*Blue*	*Green*	*Red*	*Turquoise*	*Purple*	*Grey*
2	Indian *B. indicus*	*B. taurus*					
3	Indian *B. indicus*	Eur *B. taurus*, no UK or Med	*B. taurus*, UK, Med and Africa				
4	Indian *B. indicus*	Eur *B. taurus*, no UK or Med	*B. taurus*, JE and NGu	Eur *B. taurus*, Med			
5	Indian *B. indicus*	West Eur *B. taurus*	African/NE *B. taurus*	*B. taurus*, East Med, UK	*B. taurus*, West Med, Central Eur, CH, JE		
6	Indian *B. indicus*	West Eur *B. taurus*	African/NE *B. taurus*	*B. taurus*, East Med, UK	*B. taurus*, West Med, Central Eur,	*B. taurus*, JE, CH	
7	Indian *B. indicus*	West Eur *B. taurus*	African *B. taurus*	*B. taurus* UK	*B. taurus* East Med	B taurus, West Med, Eur,	NE *B. taurus*

Table 5.2c *Interpretation of clusters produced by Bayesian admixture analysis of Dataset 3. Data is presented in* ***Figures 5.6 and 5.7****. New abbreviations used are WA=West African, UK=United Kingdom, Med=Mediterranean, Eur=Europe. Other abbreviations are cattle breed codes, which can be found in Appendix B.*

Figure 5.6 *DISTRUCT plot showing individual admixture levels from STRUCTURE analysis of Dataset 3 with the number of inferred populations set from two to seven. Med = Mediterranean breeds.*

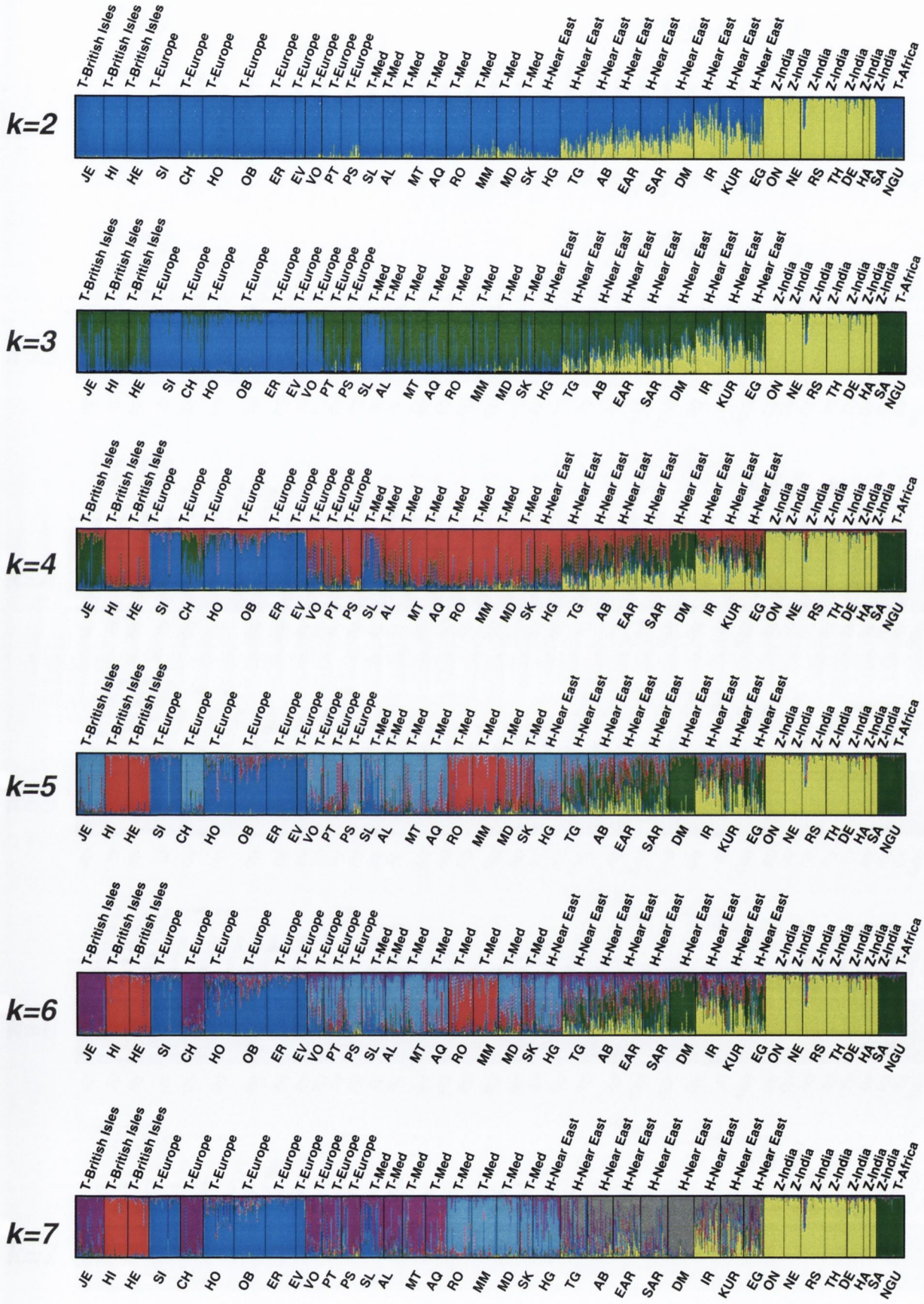
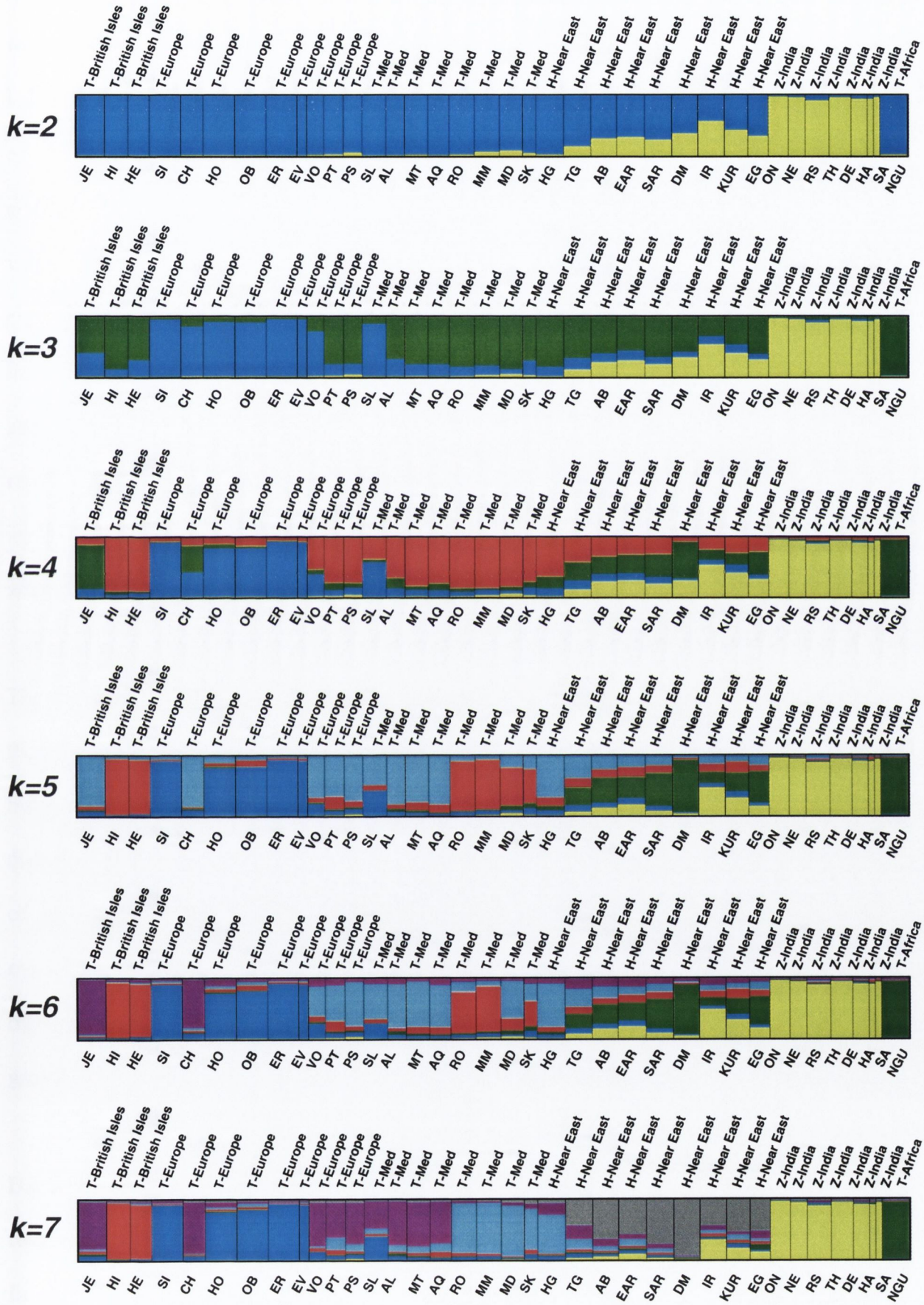


Figure 5.7 *DISTRUCT plot showing breed admixture levels from* STRUCTURE *analysis of Dataset 3 with the number of inferred populations set from two to seven. Med=Mediterranean breeds.*



5.3.1.d High values of k

The clustering analysis was carried out on each of the datasets for a high value of k (k=15 for dataset 1, k=20 for datasets 2 and 3), and results are shown in **Figure 5.8**. These high k values were only appraised for breed, and not individual admixture. While the probability of these high values is greater than for any other the other models, this is probably an artefact of the increased number of degrees of freedom in a model based on a higher k value (**Figure 5.1).** The authors of the program point out that the minimum k value that explains the structure in the data is likely to be most informative, especially if the increment between the probabilities of two k values is small. For subsequent combination analysis a k value of 3 was used. The higher probability of elevated values of k may be the result of small (and possibly insignificant) increases as the value of k increases incrementally up to the higher value, or the increased genuine structure at the breed level, which is only revealed when the data is greatly subdivided.

The first obvious feature of the analysis is the difference between the hybrid breeds and those breeds with little or no history of admixture. Most of the African and Near Eastern hybrid breeds are not categorised as discrete units with signature allele frequency distributions. Some groups of African breeds may be considered to have distinct patterns of allele frequencies and be classed as a discrete genetic entity. However, there is no evidence to support significant genetic structure among the majority of breeds with a history of admixture. In contrast, most of the non-introgressed European and Indian breeds group alone, or with one or two of their geographic neighbours.

Dataset 1 shows no divisions within or between European breeds (**Figure 5.8**). The other two plots show Jersey as a discrete unit, separate from the other United Kingdom breeds. In dataset 2, the endangered Kerry breed is also isolated from the rest of the European

breeds. The three continental European breeds in this group (CH, FR and SI) show the same grouping pattern. Dataset 3 separates most of the European continental breeds from each other, and from the other regional groups. Exceptions to this are the closely related Evolenard and Eringer breeds from Switzerland, a mainland Italian and a Greek breed (MM, SK) and three Portuguese breeds (AL, MT, AQ), which form homogeneous clusters.

A similar pattern to the one seen in Europe is also noticeable for Indian breeds which either group as one entity (Dataset 2), or as discernible units. While the Ongole and Nellore breeds are strongly grouped by one ancestry coefficient in dataset 1, one of the smaller components of their ancestry is the predominant assignment for the Sahiwal breed. Interesting, the Sahiwal cattle in dataset 1 actually consisted of imported individuals that were sampled in East Africa. The most differentiation within the *B. indicus* group is evident in dataset 3. Two main groups are evident. Once again, the southern Indian Ongole and Nellore breeds are assigned to the same group while the alternative group contains the central and northern Indian breeds Red Sindhi, Desi, Hariana and Sahiwal. Tharparker is excluded from both of these clusters.

Most of the African breeds are not strongly assigned to one group, however, the West African N'Dama breed is completely singled out by the analysis on Datasets 1 and 3. Most of the African breeds in Dataset 2 and 3 are only assigned to a single group with probabilities of about 50% and most of the inferred groups are represented within each breed group. Nevertheless, breeds from the same geographic regions or with shared ancestry appear to have similar patterns of assignment (e.g. the N'Dama samples from different countries in Dataset 2, the Ethiopian/Eritrean breeds in Dataset 1). Exceptions include the island-bound Zebu Malagasy breed of Madagascar (Dataset 1), which is

robustly assigned to a group to which no other breeds are assigned to any degree, and Pedi, a South African breed which has a high level of ancestry of a component which is found at lower levels in other southern African samples.

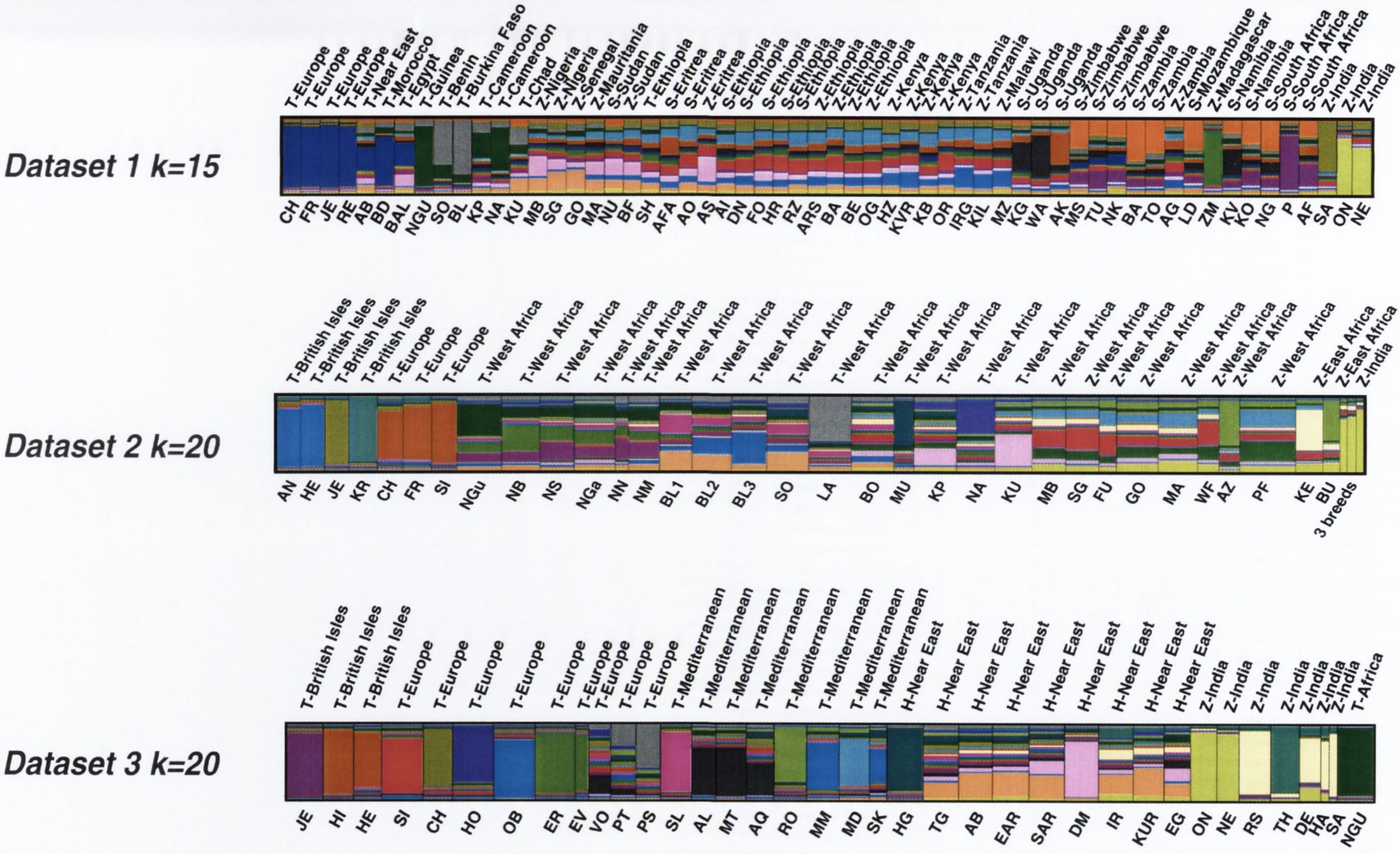


Figure 5.8 *DISTRUCT diagram of 3 datasets with k set to 15, 20 and 20.*

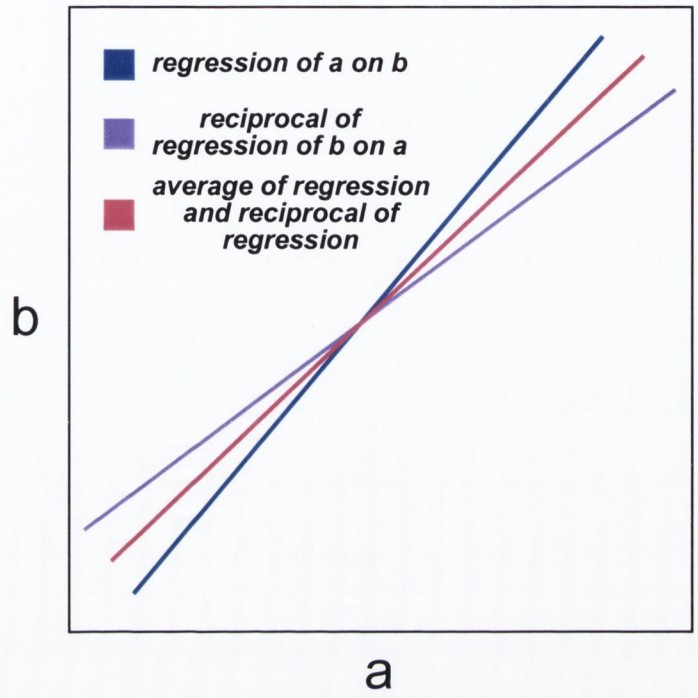
5.3.2 COMBINED DATASETS

5.3.2.a Combining Data

Genetic diversity and admixture levels in these populations have been discussed in previous chapters and elsewhere (EDWARDS *et al.* 2000; HANOTTE *et al.* 2002; LOFTUS *et al.* 1999; MEGHEN *et al.* 1994). This chapter utilises transformation methods to incorporate information from the different sets of microsatellite markers and produces a global analysis of population parameters.

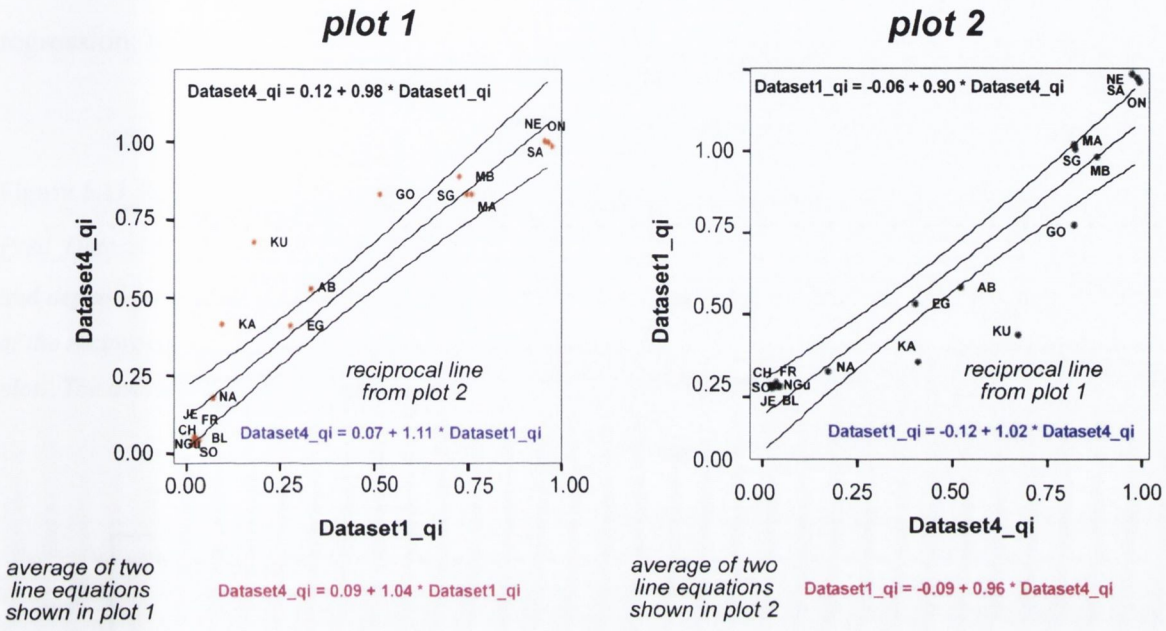
Least squared distances are traditionally minimised on the vertical axis of linear regression when one of the variables is considered to be independent. However, if there is error in the data, the regression of any variable (**a**) a on any other variable (**b**) will not be an exact reciprocal of the regression of variable **b** on variable **a**. The optimum line for transforming between these two variables is the average of these two regression lines. A schematic representation of this is shown in **Figure 5.9**. The transformation method outlined in Section 5.2.3 is illustrated below for the admixture coefficient q_i and Datasets 1 and 4. As a general rule, it is necessary to have a significant correlation between the measures of the variable in the two datasets. The correlation parameters for q_i are shown in **Table 5.4**. As there is no real independent and dependent variable, and there is an error associated with both, the regression is carried out using each dataset as both predictor and dependent.

Figure 5.9 *Schematic showing the average line (pink) produced from the regression of variable* ***a*** *on variable* ***b*** *(blue), and the reciprocal of the regression of variable* ***b*** *on variable* ***a*** *(purple). If there were no errors associated with the data, the blue and purple lines would be identical. This is rarely the case with real data, therefore the average line will best represent the relationship between variable* ***a*** *and* ***b****.*



Plots 1 and 2 (**Figure 5.10**) show the regression plots for q_i values for common breeds in Datasets 1 and 4. In plot 1, Dataset 1 is treated as the independent variable, while Dataset 4 is the independent variable in plot 2. If there were no errors associated with the data, the reciprocal of the regression line in plot 2 would be identical to the regression line in plot 1. The equation of the reciprocal of the plot 2 regression line is shown in the lower right hand corner of plot 1 (purple). It is not the same as the original regression function (black), which is shown in the top right hand corner of plot 1. Therefore the average of the two line equations (pink) is calculated and used for the transformation. The equivalent functions are shown for the second regression in plot 2.

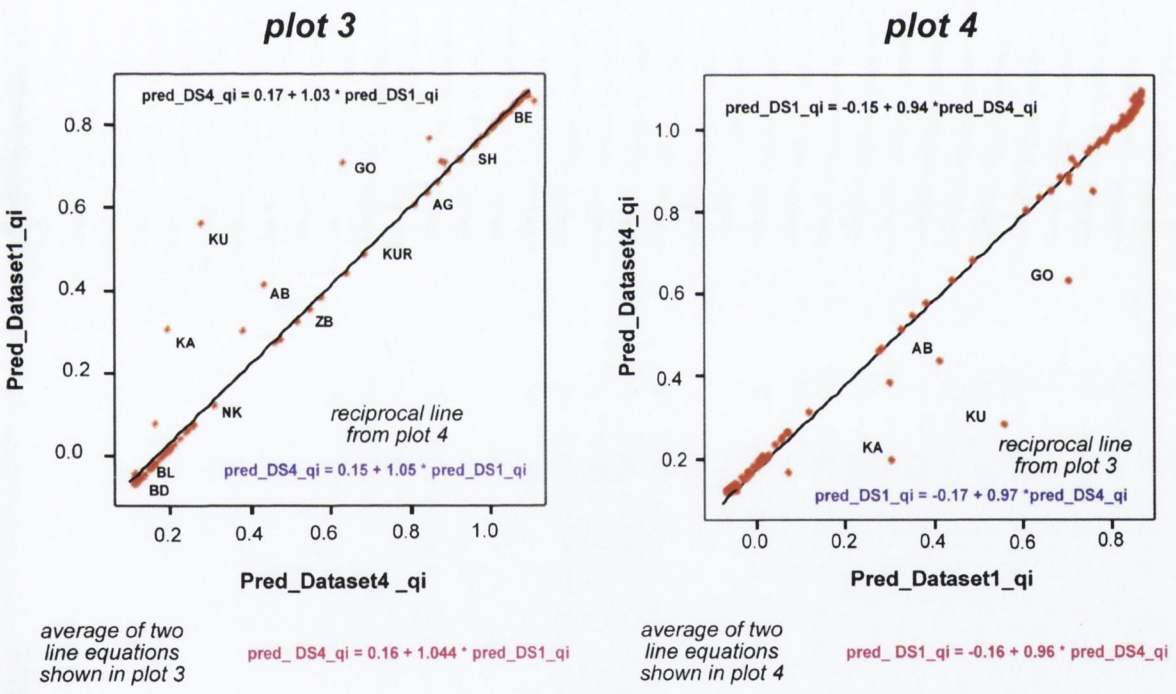
Figure 5.10 *Regression plots of q_i values for common breeds found in Datasets 1 and 4. The regression is shown with each dataset acting as both independent and dependant variable. Regression equations are shown in the top left hand corner of plots. The reciprocal line equation (purple) from the opposite regression is shown in the lower right hand corner of each plot. The average of the two equations (pink) is shown below the plots. 95% confidence intervals are shown around the regression line.*



The average equations are used to transform each value of the relevant datasets to a predicted value, i.e. the equation below plot 1 is used to transform values from Dataset 1 to a new dataset of q_i values denominated predicted_Dataset1_***qi***, and the equation below plot 2 is used to transform values from Dataset 4 to a new dataset of q_i denominated predicted_Dataset4_***qi***. For breeds that are not included in one or other of the datasets, the constant in the equation is added to the original value from the alternative dataset. Every breed, not just those held in common between the two datasets, is now included in a second regression analysis. This second transformation step is necessary, as the predicted dataset has two predicted values for the common breeds, but one predicted and one original value for the breeds that occur in only one dataset. A second transformation ultimately allows averaging of two predicted values for each breed, a common treatment that maintains the relative position of each breed in the final dataset. Thus, the process outlined above is

repeated for these two new datasets and the regression plots are shown in **Figure 5.11.** While the second set of regression lines (**Figure 5.11**) are not very different from the first (**Figure 5.10**), the associated error is reduced in the second set of regressions. Outliers that were present in the first regression analysis do not necessarily fit better in the second regression, but the overall fit (as measured by r^2 (adj), **Tables 5.3 and 5.4**) is improved.

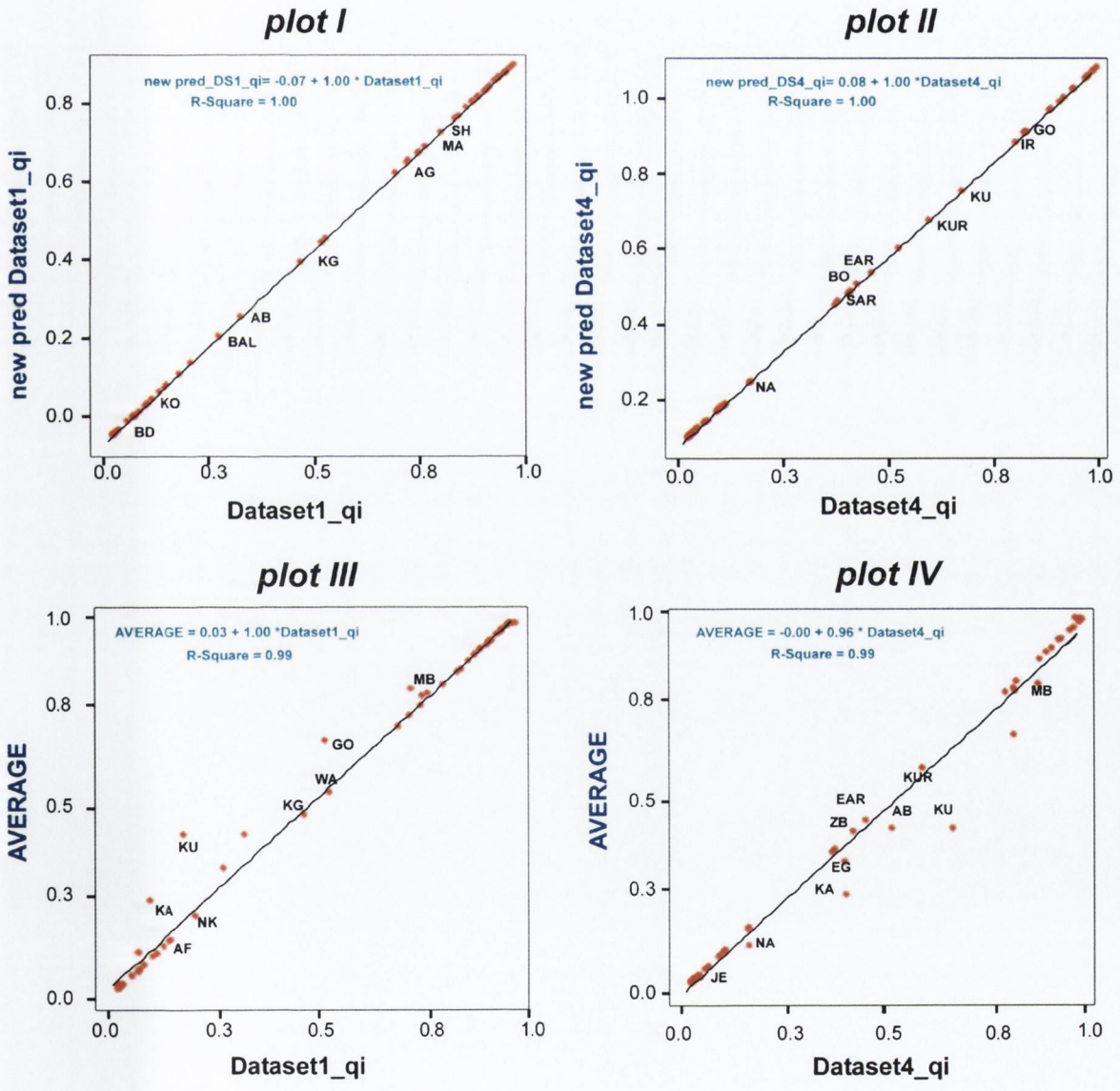
Figure 5.11 *Regression plots of q_i values for common breeds found in Pred_Dataset1_q_i (pred_DS1_ q_i) and Pred_Dataset4_ q_i (pred_DS4_ q_i). The regression is shown with each dataset acting as both independent and dependant variable. Regression functions are shown in the top left hand corner of plots. The equation of the reciprocal line (purple) from the alternative regression is shown in the lower right hand corner of each plot. The average of the two functions (pink) is shown below the plots. The outlier breed codes are given.*



The average of the pair of regression equations is finally used to transform each of the first predicted datasets. The final transformation includes all breeds, and the values produced (denominated *new predicted* values) from each dataset are averaged for each breed to obtain a final estimate of the q_i admixture coefficient, or another variable in question.

To examine the relationship between the original datasets and their finally transformed counterparts (new predicted), regressions were carried out to ensure that point estimates in the two dataset had retained their integrity. These regressions confirm this, and are shown as plots I and II in **Figure 5.12**.

Figure 5.12 *Scatterplots of regressions carried out between raw and transformed datasets. Plots I and II clearly show the retained correlation between each of the original datasets and their (twice) transformed counterparts. Plots III and IV show the slightly less perfect correlation between the original datasets and the final averaged dataset which was used in subsequent analyses.*

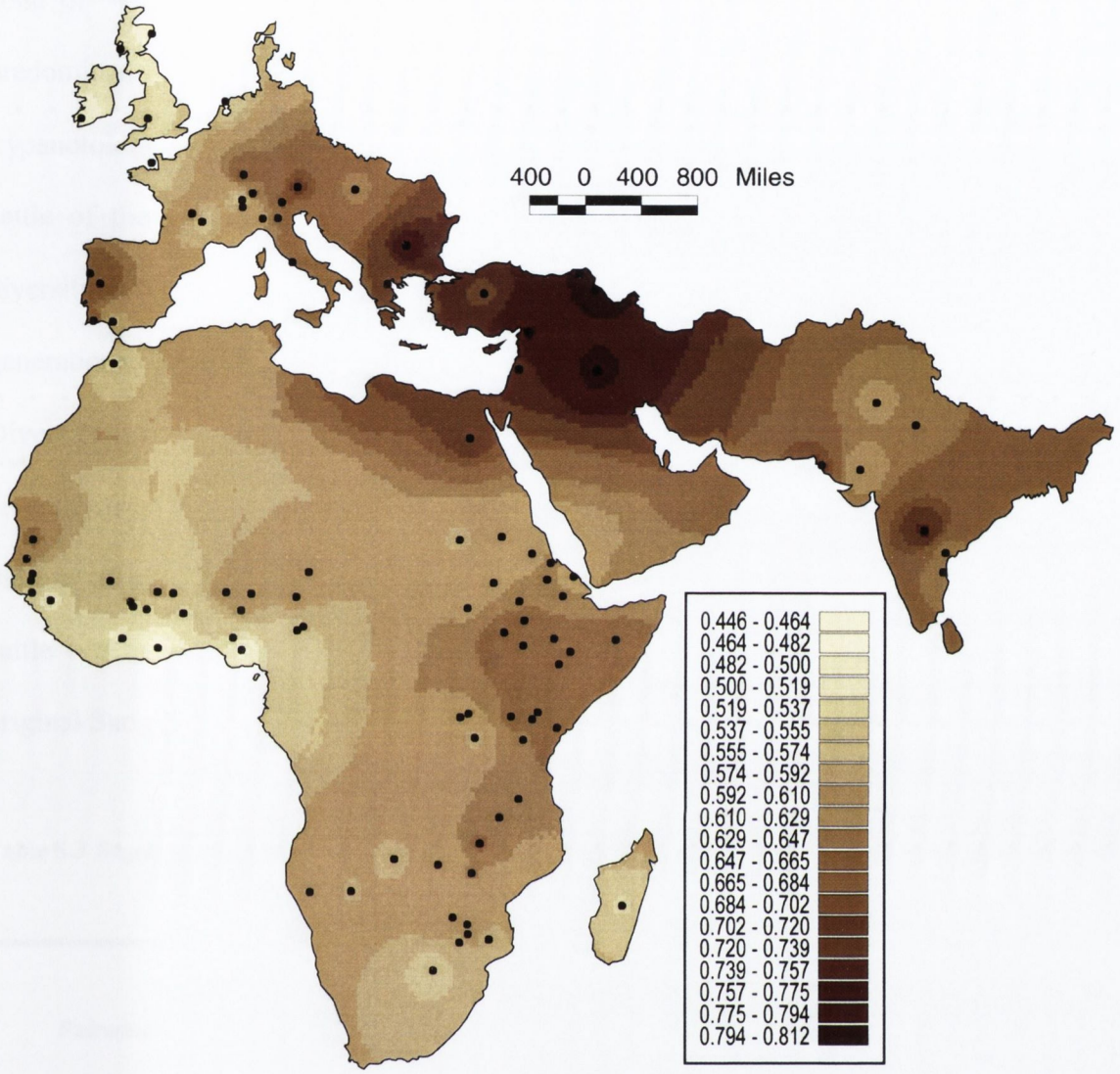


Ultimately, the relationship between the original datasets and the final averaged values is most critical and these are shown in plots III and IV (**Figure 5.12**). The Kuri breed continues to be an outlier, a position that has been maintained since the original regression analysis. However, the slopes and highly significant r-squared values suggest that the within-dataset values from each initial dataset are retained in the final average dataset, which contains all breeds from both datasets.

5.3.2.b Unbiased Heterozygosity ($\hat{H}$)

$\hat{H}$ is the expected probability that an individual will be heterozygous over the assayed loci in a multi-locus system. Regression parameters (p <0.0001, r^2(adj)= 0.826 (regression 1) and 0.988 (regression 2)) for both the first and second regressions using datasets 1 and 2 were highly significant and are shown in **Table 5.3**. When the combined dataset (1&2) was compared to dataset 3, the correlation between the two datasets was not quite as significant (**Table 5.3**) but still enabled combination. This is most likely due to the fact that there are fewer breeds in common in the second set of regressions (n=11), and that the range of values for heterozygosity tends to be limited. However, in both cases, the r^2 values are increased each time the data are transformed to a best-fit line. This demonstrates that prediction power is enhanced and noise reduced as the two datasets converge after transformation.

Figure 5.13 *Contour map of average predicted* $\hat{H}$



A synthetic map of combined predicted $\hat{H}$ values is shown in **Figure 5.13**. Highest predicted values are found in the Near East, followed by India and East Africa. The Near East is an important centre of domestication so it is not surprising that it should contain high levels of diversity. The cattle of Europe have gradually less diversity as distance from the Near East increases, and smaller subsets of the original diversity remain. Heterozygosity levels in the Near East have also been increased by the arrival of *B. indicus* cattle from the Indo-Pakistan region (LOFTUS *et al.* 1999). The Desi breed in India has a very high level of heterozygosity but this can be explained by its origins, as it is not an

authentic breed (see Section 2.2). Low levels of heterozygosity are present in the tsetse zone of West Africa where trypanosomiasis is endemic. The cattle of this area are predominantly indigenous African taurine breeds, such as N'Dama, which are trypanotolerant. These cattle have undergone strong selection for this trait. The local cattle of the Britain and Ireland, such as Jersey, Highland and Kerry also show low diversity according to this measure. These local breeds may have been isolated for many generations, and show even lower heterozygosity levels than most West African breeds. Diversity levels in Europe tend to be low as the breeds of the region are a subset of the diversity from the Near East and many have been subject to selective breeding, including the use of artificial insemination. High diversity levels can be seen in some Portuguese cattle breeds while the Afrikaner cattle of South Africa, which are derived from the original Sanga cattle (CLUTTON-BROCK 1989), exhibit a relatively low level of diversity.

Table 5.3 *Regression parameters for transformation of diversity parameters*

Pairwise linear regression between:		$\hat{H}$ *(p)*	$\hat{H}$ r^2*(adj)*	$\bar{A}$ *(p)*	$\bar{A}$ r^2*(adj)*
Data1	***Data2***	0.000	0.826	0.000	0.708
Data1 (pred)	***Data2 (pred)***	0.000	0.998	0.000	0.824
Data 1&2 (combined)	***Data3***	0.002	0.582	0.001	0.644
Data 1&2 (pred)	***Data3 (pred)***	0.000	0.626	0.000	0.730

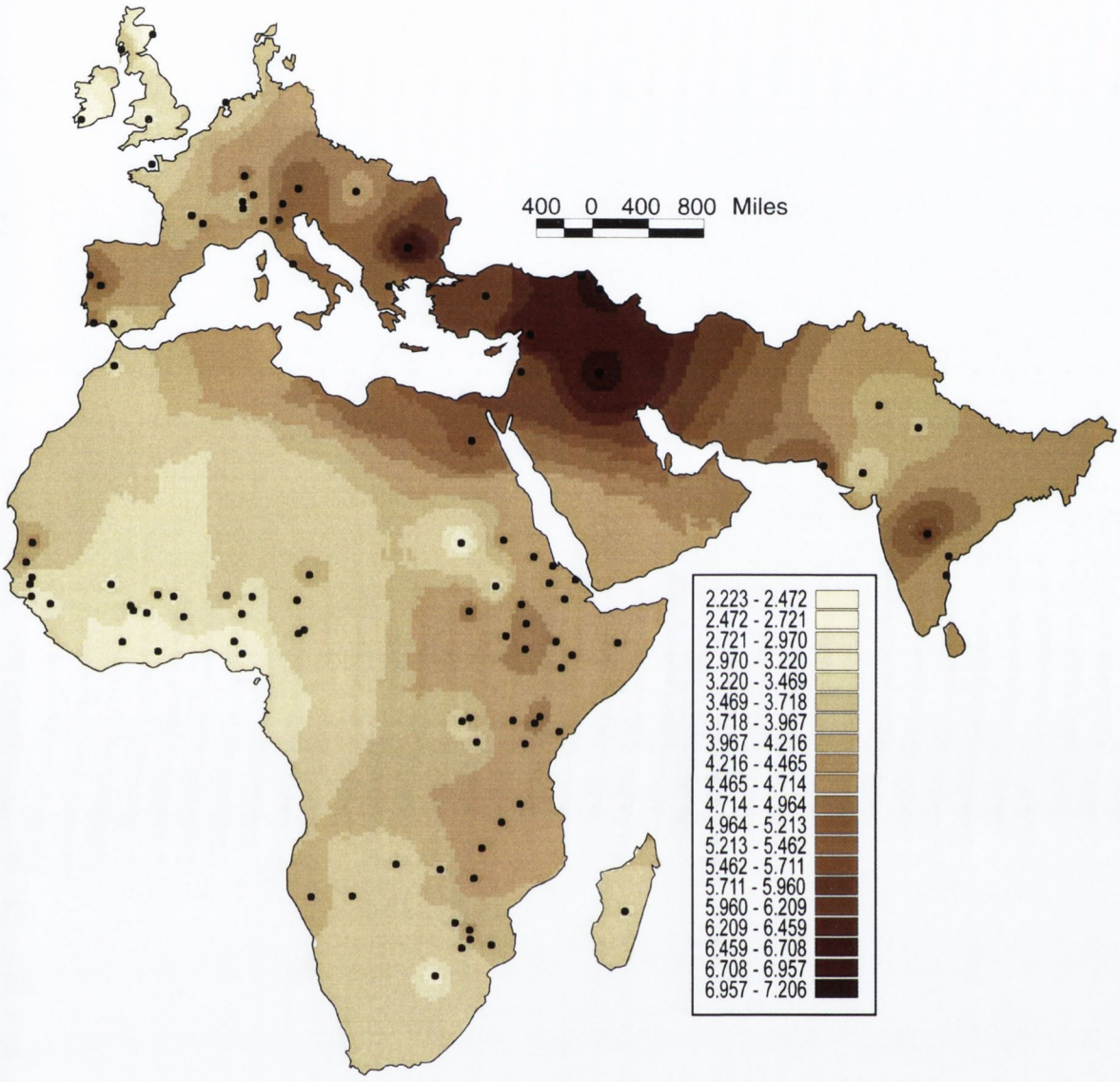
5.3.2.c Mean number of alleles ($\bar{A}$)

$\bar{A}$ is a measure that is highly sensitive to admixture and can be especially useful in identifying recently admixed populations. Values from three datasets were combined using regression analyses. Regression parameters for combining datasets 1 and 2 were

highly significant. *p*-values and r^2 (adj) values for the regressions between datasets 1 and 2 and between datasets 1&2 (combined) and 4 are shown in **Table 5.3**. The *p*-values (<0.0001) for the two regressions show a highly significant correlation between the two datasets (**Table 5.3**). The high values of the conservative measure, r^2 (adj), show that there is prediction ability and limited noise in the data.

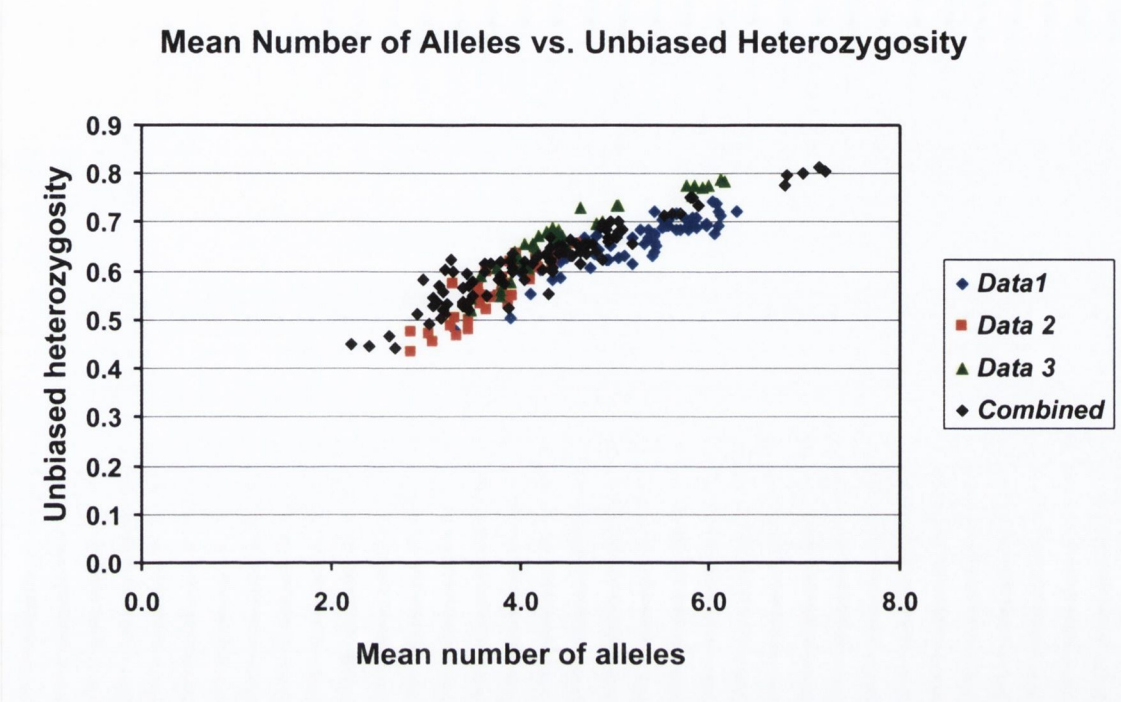
The global pattern of $\overline{A}$ (**Figure 5.14**) is very similar to unbiased heterozygosity (**Figure 5.13**). Once again it is evident that highest diversity levels are found in the Near Eastern breeds, which may have incorporated substantial diversity from the progenitor wild population, but have also received a large genetic input from Indian domestic cattle (LOFTUS *et al.* 1999). Levels of diversity are quite high in North India, another putative centre of domestication and admixture. Within the African continent, it is not the West African hybrid breeds that have the highest number of alleles, but the Egyptian and East African samples.

Figure 5. 14 *Contour map of size-adjusted mean number of alleles ($\overline{A}$)*



Multiple transformations should not change the relationships between or within the datasets. In the original datasets there is a positive linear relationship between the two calculated diversity measures. This relationship is maintained in the combined dataset **(Figure 5.15)**.

Figure 5.15 *Scatterplots showing the relationship between* $\overline{A}$ *and* $\hat{H}$ *in Datasets 1, 2 and 3 and combined datasets.*



5.3.2.d Admixture

Three admixture coefficients were estimated, following examination of likelihood values for models inferring different numbers of parental populations. Each admixture coefficient was characterised by populations in which it was inferred to be close to one. The three coefficients are designated q_a (African taurine), q_e (European/Near Eastern taurine) and q_i (Indian zebu). Datasets 2 and 3 were pooled for the seven loci they hold in common and are denoted dataset 4 for the purposes of the admixture analysis. Despite being comprised of only seven markers a highly significant result is obtained for the regressions between this combined dataset and Dataset 1, which is comprised of 15 loci. While it would be ideal to have more loci in dataset 4 to improve the accuracy of our estimates the linear transformation combination method can still be applied, as there are highly significant correlations between the two datasets. The admixture results were combined for each of

the three admixture coefficients calculated for the inferred populations (q_e, q_a and q_i). The regression parameters are shown in **Table 5.4**. There is a highly significant positive linear relationship between each of the admixture parameters in the two datasets.

Table 5.4 *Regression statistics for pairwise regression analyses carried out to combine datasets. The table shows the first and second pairwise regressions between datasets 1 and 4 for each of the 3 admixture components.*

Pairwise linear regression between:		q_e ***(p)***	q_e r^2***(adj)***	q_i ***(p)***	q_i r^2***(adj)***	q_a ***(p)***	q_a r^2***(adj)***
Data1	***Data4***	0.000	0.973	0.000	0.869	0.000	0.781
Data1 (pred)	***Data4 (pred)***	0.000	0.996	0.000	0.973	0.000	0.958

Synthetic contour maps for each admixture parameter are shown in **Figures 5.16, 5.17 and 5.18**. The admixture coefficient that is characterised by an ancestral African *B. taurus* genetic component (q_a) is shown in **Figure 5.16**. Here it is clear that the West and Southern African breeds are the ones that have largely escaped introgression. In West Africa the proportion of this component in a breed can be related to the tsetse challenge in its locality, as the indigenous cattle can tolerate the trypanosomes carried by this biting fly. The cattle of East Africa are almost completely indicine in character. It is noteworthy that a fraction of African influence is observed to extend into Europe (via Iberia) and the Near East.

Figure 5.16 *Contour map of new predicted* q_a

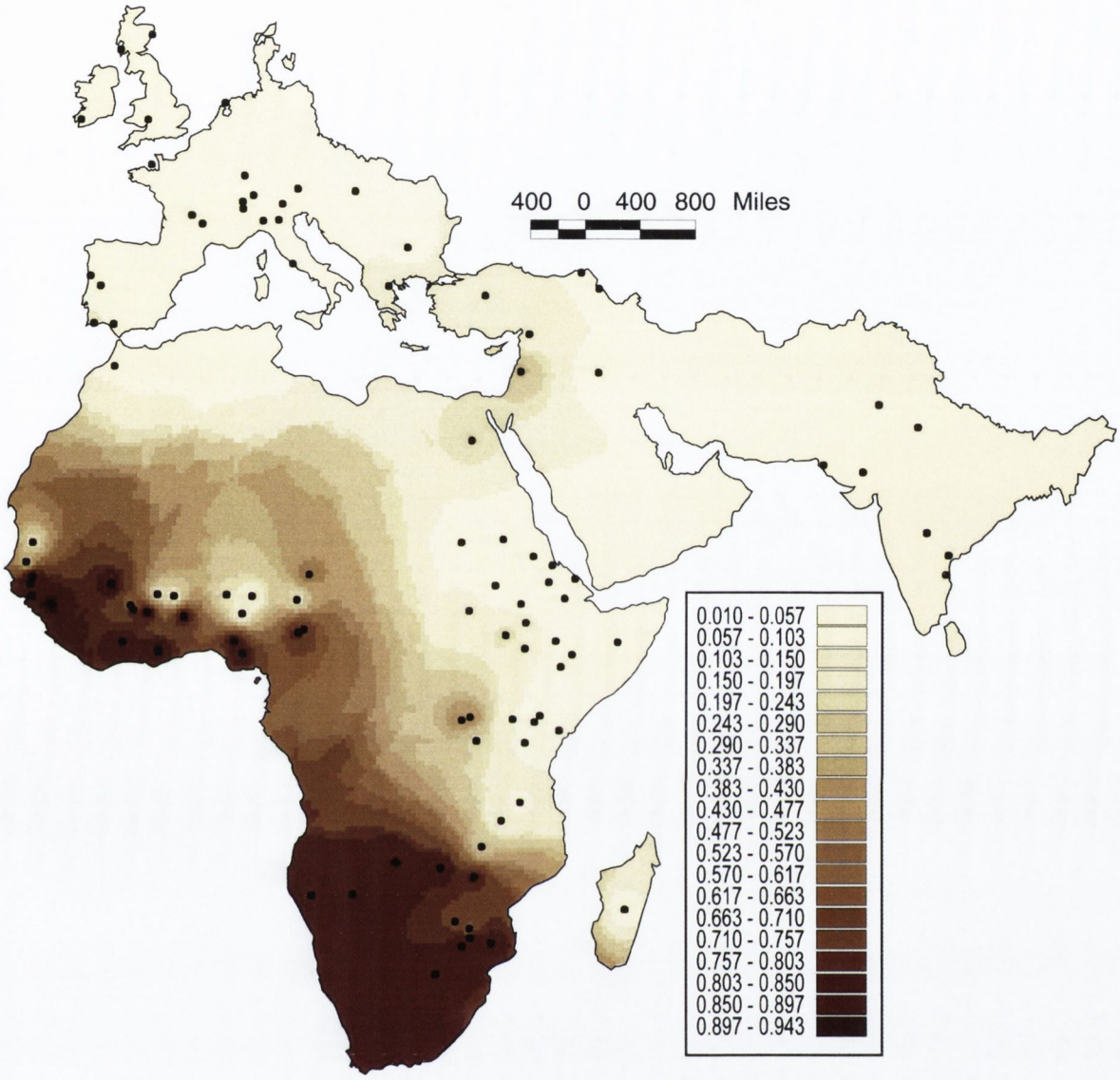
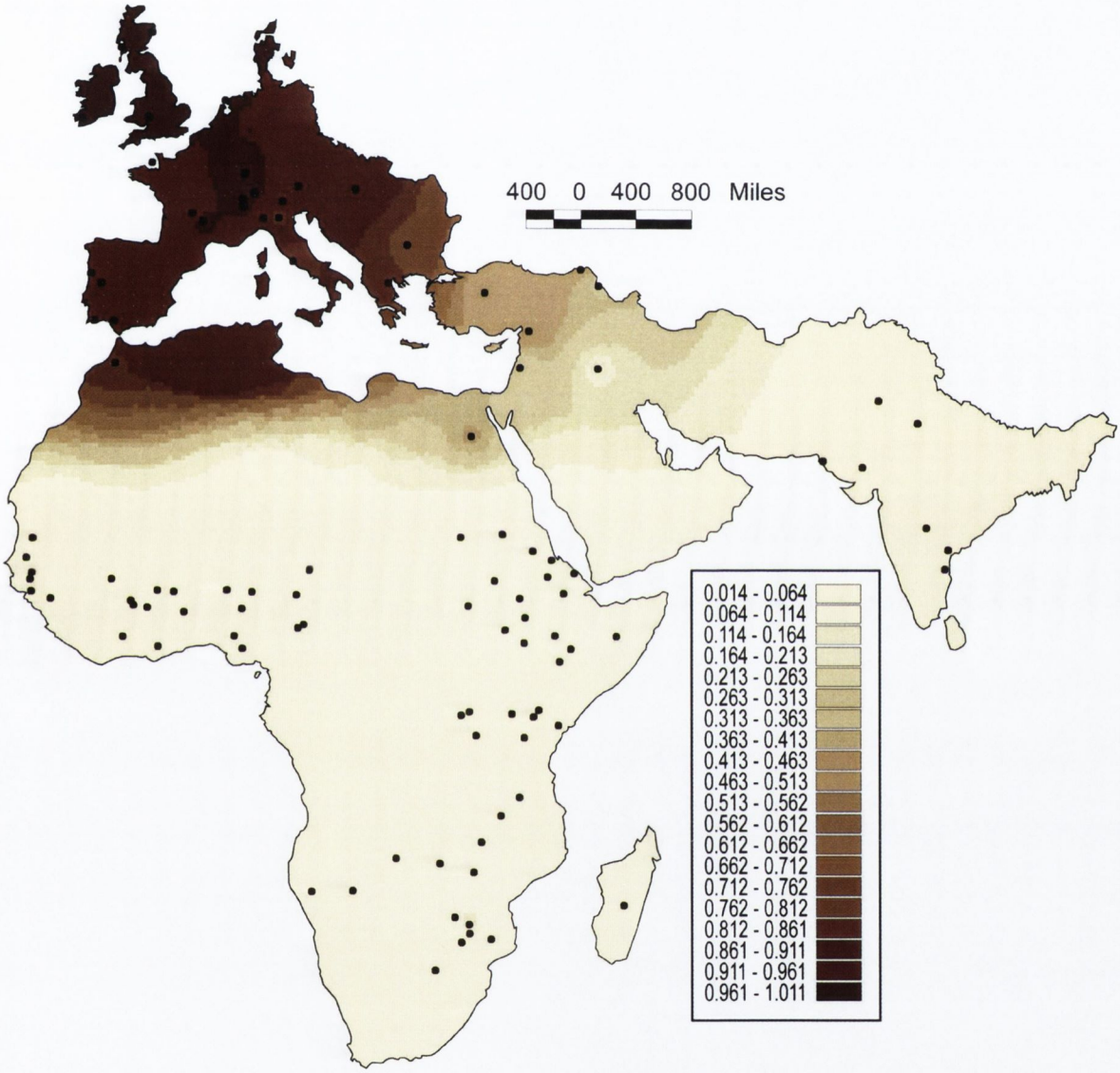


Figure 5.17 shows the admixture coefficient q_e that represents the ancestral influence of European and Near Eastern *B. taurus*. Cattle found at this time in the Near East have been influenced by significant amounts of introgression from further East and would all be classified as admixed populations. European cattle represent a subset of these originally domesticated *B. taurus*. It is possible to see a clinal reduction in taurine influence from west to east across the Near East, where incursion of zebu cattle has had a significant

impact (LOFTUS *et al.* 1999). Moroccan Blonde and Egyptian cattle show considerable input from the Near Eastern/European genetic pool.

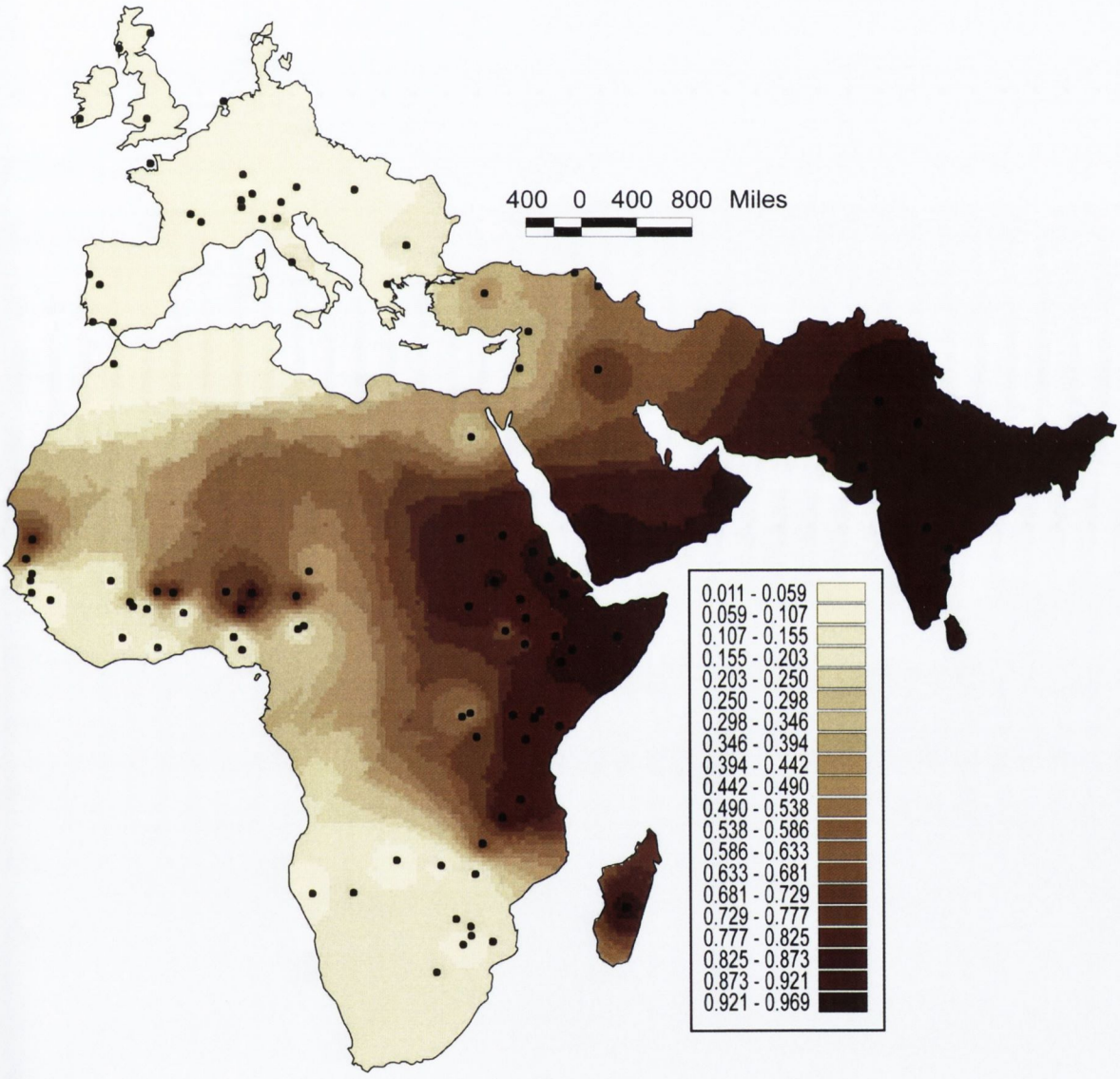
Figure 5.17 *Contour map of new predicted* q_e



The final major axis of admixture is the result of mass movement of *B. indicus*, dry adapted- humped cattle, from Arabia and the Indo-Pakistan region, where they were domesticated, to the increasingly dry areas of the Middle East and East Africa. This is represented in **Figure 5.18** by the coefficient q_i. The cattle of Madagascar and the Horn of

Africa are almost indistinguishable from the present-day representatives of the original contributing indicine cattle. Hanotte and colleagues also observed this phenomenon (HANOTTE *et al.* 2002).

Figure 5.18 *Contour map of new predicted* q_i



The gradient of zebu influence through the Near East can be seen to stretch from East to West, mirroring the decrease in taurine character. This implies that cattle also moved along landward routes from India and Pakistan

5.4 DISCUSSION

The aim of this study was to provide a global view of diversity and admixture by a) comparing structuring within three extensive datasets and b) transforming derived variables to produce a single dataset of diversity and admixture measures.

5.4.1 POPULATION STRUCTURE

An examination of the results of the Bayesian admixture analysis carried out on the separate datasets reveals some consistent trends among them. In previous studies, each increase in k in the Bayesian admixture analysis tended to split an existing cluster and its individuals were assigned with different levels of confidence to one of the two new clusters, proceeding in a hierarchical succession of bifurcations (ROSENBERG *et al.* 2002). However, this is not always the case with the data in the current analysis. An initial inspection of the overall pattern as k increases shows that the clustering that occurs in Europe and India is relatively clear, but that the African populations form more complex and intricate groups. Unsurprisingly, the main division that is seen within each of the three sets of microsatellite marker genotypes is that between *B. taurus* and Indian *B. indicus*. Indian zebu breeds, defined by the yellow cluster in each of the figures, are maintained as a discrete unit, when k is greater than three. As the value of k increases, some breeds are demarcated as distinct units, while others show ancestry from several clusters.

According to the likelihoods associated with each k value (**Figure 5.1**), the most appropriate estimates for k are 3, 3 and 4 for datasets 1, 2 and 3 respectively, and these merit particular attention. In Dataset 1, at k=3, there is clear division of the *B. indicus* element from two taurine quantities. The zebu incursion into the African continent is evident and interestingly, West and North African taurine breeds group with European

breeds, and not with Southern African *B. taurus*. However, examination of subsequent values of k demonstrates that this cluster is unstable and is not preserved for most of these African breeds. It is noteworthy that it is the North African breeds from Egypt and Morocco that continue to group with Europe, a reflection of some reciprocal gene transfer, especially a contribution of European and Near Eastern influence into these areas.

The most informative model for dataset 2 (k=3) shows clear segregation of the *B. indicus* group, and two *B. taurus* clusters separating Europe and the Near East from Africa. At higher values of k, partitioning increases and produces clusters with some geographical congruence. At k=7, the Bayesian grouping of West African taurine breeds corresponds well with their physical location in the region. The N'Dama and Baoule samples from different countries group together, as do the taurine breeds sampled in Benin (Somba and Laguna). Breeds from the eastern Nigeria and Cameroon (Namchi, Muturu and Kapsiki) form a distinguishable cluster. Indicine-type breeds do not tend to form discrete units in either dataset 1 or 2, probably as a result of their recent introduction to the continent.

Within Europe there is little or no differentiation visible according to the groupings in datasets 1 and 2 (**Figures 5.3 and 5.5**). In dataset 2 at k=5, Aberdeen Angus, Jersey and Kerry appear to form a discrete group but are once again strongly clustered with the other European breeds when the value of k is increased by one. This result is difficult to interpret. Previous analysis using some of the data from this dataset found strong breed clustering, particularly for the Swiss Simmental breed and breeds originating from the British Isles (MACHUGH *et al.* 1998).

Dataset 3 differs from the other two, as a k value of 4 appears to be optimum. This third dataset contains the largest number of European breeds and also many breeds from their region of origin, the Near East. At k=4 there is a (subsequently unstable) grouping of Jersey and N'Dama, which is not authentic and disappears at higher k values. Contrary to the previous two datasets, there are several clear groupings of European breeds (**Figure 5.7**). The emergence of a North Central European component comprised of the breeds from Switzerland, Germany and the Netherlands (blue) is evident, and separate from a component that is shared by Britain, the Mediterranean and Near East (red).

Dataset 3 reveals other interesting groupings within Europe and the Near East when the value of k is increased. At k=5, the European breeds can be broadly divided into three groups. The North Central group remains intact, while Highland and Hereford cluster with Italian breeds, to the exclusion of the other British island breed, Jersey, which actually seems to be more closely related to the French and Iberian breeds. The latter relationship is maintained at k=7, but the eastern Mediterranean grouping is not very stable and is subsequently sub-divided, with the two British breeds grouping separately. However it does mirror results from Y-chromosome data, where high diversity levels in Britain support the idea of a dual stream of genetic input to north-western Europe; through the centre of Europe along the Danubian route, and directly from Iberia (EDWARDS 2003; THOMAS 1996).

Archaeological evidence does not support a significant time differentiation between the arrival of the Neolithic and agriculture in the eastern and western areas of the Mediterranean (HASSAN 2000; ZILHAO 2001). Therefore, the differentiation that is observable between the eastern and western Mediterranean breeds (**Figure 5.7**) may

merely be the result of breed separation as cattle spread along a littoral route from the centre of domestication in the Near East (TROY *et al.* 2001). The high levels of diversity in Portugal (**Figure 5.13 and 5.14**) supports previous findings that the cattle that originally entered Iberia along a contemporary maritime route (RENFREW 1996; ZILHAO 2001) have experienced introgression from Africa (CYMBRON *et al.* 1999). It is also possible that genetic structuring within European cattle is the result of modern breeding practices, especially in those breeds that have been selected for commercially important traits. It is interesting that there is a greater tendency for individual breed partitioning in the northwest Europe than in the Mediterranean and the Near East (**Figure 5.8**). This may be a consequence of the more rapid transfer of cattle along the coastal route and subsequent communication between Mediterranean breeds.

The genetic structure that is displayed in the Near East (Dataset 4) is highly consistent with previously published data, which revealed a cline of *B. indicus* influence from East to West. Interestingly, this zebu component remains largely intact as the value of k is increased, while the taurine component is incrementally subdivided. Each of the new taurine elements is also found at significant levels in European breeds, perhaps a signature of their derivation from the Near East. At k=7 a characteristic Near Eastern clustering is present, but there is a lack of individual breed structuring.

On the other hand, previous studies have found genetic differentiation between groups of cattle breeds within Europe using other systems. A study analysing 86 genes coding for blood and milk proteins grouped what were denominated Balkan breeds, Balkan-Alp breeds and then synthetic breeds that had a history strong artificial selection (MEDJUGORAC

et al. 1994). Kantanen and colleagues (2000) used autosomal microsatellites to find significant divisions between indigenous Nordic breeds (KANTANEN *et al.* 2000).

Deciphering the structuring within the African continent is a more challenging task. For all the datasets, African *B. taurus* breeds, especially Guinean N'Dama, are strongly allocated to a single inferred group, suggesting a distinctive allele frequency pattern within this group. Presumably these breeds have been isolated both by distance and their presence in the tsetse zones. Dataset 1 reveals a clear distinction between Western and Southern African taurine breeds, which may be the result of the considerable time that these two groups have been geographically separated (**Chapter 4**). The continual separation of southern African taurine breeds from their West African counterparts through higher values of k implies that recent gene flow between the two areas has been limited, and that some of the earliest domesticates in the south may already have been genetically differentiated by the time they migrated to the Cape. It is generally accepted that that there must have been dispersal of cattle from a north African domestication centre to the south and west, during which differentiation may have been caused by the colonisation process and population bottlenecks due to disease (GIFFORD-GONZALEZ 2000; HANOTTE *et al.* 2002).

It is noteworthy that both African components (represented by blue and green in k=3 model, **Figure 5.3**) are present in the Baladi cattle of Egypt at approximately equal levels. This is especially interesting, because it is what would be expected if these two elements of variation were originally present within the early African early domesticated cattle and represent genetic differentiation within the African taurine from a single ancient centre of origin, close to Baladi cattle are found today (HANOTTE *et al.* 2002; MACDONALD 2000).

Anatolian Black cattle from the Near East show a similar pattern of diversity within the taurine element of variation.

The STRUCTURE analysis seems to show several multiple elements of indicine genetic variation in Africa, rather than geographically discrete indicine groups, as is the case with the *B. taurus* components. The Africa zebu breeds in Datasets 1 and 2 have multiple layers of *B. indicus* variation but show little between breed differentiations (**Figures 5.3 and 5.5**). The non-structured pattern of zebu, as compared to *B. taurus* breeds, is almost certainly related to the later arrival of zebu into the African continent. *B. indicus* cattle entered Africa from several different regions; initially via the Suez (EPSTEIN 1971), then from the Arabian peninsula and via coastal routes to the Horn of Africa (HANOTTE *et al.* 2002). The major introduction of zebu is likely to be recent; therefore there has been limited time for differentiation to occur. The lack of clustering of even geographically proximal *B. indicus* breeds is almost certainly a hallmark of their late arrival on the continent.

It is clear that the major regional groupings of cattle in India, Europe and Africa are supported by lower k values. However the high k values (**Figure 5.8**) give us an insight into whether *a priori* grouping of individuals into breed units and regional units is a reflection of the true genetic make up of the populations. It is directly observable that the analysis reveals broad structural differences between hybrid and non-introgressed breeds in all datasets. European, Indian and some of the West African breeds are shown as discrete distinguishable units, while hybrid breeds from Africa and the Near East have multiple ancestry coefficients. Genuine breed structure seems to be revealed, as the Indian breeds Ongole and Nellore (a breed derived from Ongole stocks transported to Brazil) consistently

group together, even to the exclusion of the other Indian breeds. British breeds also cluster to the exclusion of other European breeds.

Most African breeds, especially those with histories of considerable admixture do not appear as discrete units and lack structure. This is obviously due to the dual processes of recent introductions and hybridisation, and also to breeding practices. In developing regions there may be high levels of unregulated crossbreeding which can severely complicate the task of breed classification. In these areas distinct cattle types are associated with different groups of pastoralists. It is to be expected, therefore, that the structure of cattle breeds that are maintained in a traditional livestock environment would not be as defined as those reared in developed regions of the world (MEGHEN *et al.* 1994). This will contrast strongly to breeding and husbandry practices in more developed regions of the world where practices such as studbook maintenance and artificial insemination may result in low effective population sizes and little genetic differentiation within a breed (FARNIR *et al.* 2000). In conclusion, genuine groupings appear to be maintained through different k values, with different levels of information being revealed depending on how populations are divided. Moreover, authentic clusters seem to be sub-divided further into discrete clusters as the value of k increases. Nonetheless, each k value does represent an imposed, not inherent, structure and should be treated as such.

5.4.2 GLOBAL DIVERSITY

A global picture of diversity in any species may prove to be a valuable tool when defining areas of high biodiversity for conservation or study. In any localised study there will be samples of relatively lower and higher diversity. While this information may be useful in a local situation, it may be far more valuable when placed in a global context, and this is the advantage of being able to amalgamate datasets incorporating samples from different areas of the world.

In an isolated population, heterozygosity measures will decrease with drift. Population bottlenecks, inbreeding, founder effects and selective breeding may also act to decrease heterozygosity, as alleles are lost from a population. Any event that results in gene flow between populations, such as miscegenation and introgression (especially continuous), will be inclined to increase the number of heterozygotes in a population. Typically, isolated breeds and breeds undergoing intense selection will have low $\overline{A}$ values while breeds that have been recently introgressed or which are experiencing high levels of gene flow will have high $\overline{A}$ values. Therefore areas of high diversity in **Figures 5.13 and 5.14** may either be associated with centres of domestication, or more likely, with areas where admixture has occurred. In contrast, breeds that are undergoing either natural or artificial selection or isolation would be expected to have low diversity measures.

Theoretically, it might be expected that admixture levels of 50% taurine and 50% indicine would produce the highest diversity levels. In West Africa, above the tsetse zone (**Figures 3.1 and 3.7**), there are many breeds that have taurine/indicine admixture levels of about 50 per cent, and if dataset 2 is examined in isolation these breeds do have the highest levels of diversity. However, when the datasets are combined, it appears that most of the East

African humped breeds, which are predominantly indicine, have higher levels of diversity. While this may seem initially counterintuitive, East African zebu breeds have received a large proportion of their genome from imported *B. indicus* and this admixture contributes to higher levels of genetic diversity. Three-way admixture between African and Near-Eastern *B. taurus*, and Indian *B. indicus* may also be a feature of this region. As indicine cattle spread into southern Africa and East into the Sahel region, diversity may have been lost in the same way that diversity was lost as taurine cattle spread from the Near East into Europe. The low genetic diversity levels of African taurine cattle of the West Coast countries may be due not only to selective pressures but also to the fact that they could represent a very small subset of the diversity that was present in the original African domestication event, which probably took place in North Africa with subsequent migrations to other parts of the continent (MARSHALL and HILDEBRAND 2002). The Egyptian Baladi cattle breed has both Near Eastern and African influence and this may account for the elevated number of alleles.

The diversity levels in India are intermediate relative to areas where large-scale hybridisation has taken place between the two sub-species of cattle. However, diversity levels may have been underestimated in these breeds, as most bovine microsatellites have been developed in economically important taurine breeds and may have relatively few indicine alleles. As a centre of domestication fairly high levels of genetic diversity are expected in these breeds. The highest levels of diversity in India are seen in the central Desi "breed". These animals do not represent a recognised breed and are a mix of non-descript local cattle. They may have had input from different sources explaining their relatively high diversity levels.

Within Europe, there is a cline of loss of diversity from south-east to north-west, a pattern which has been well characterised as the spread of the Neolithic from the Near East to Europe (CAVALLI-SFORZA 1996) and which supports findings of Troy and colleagues (2001), who demonstrated that the diversity present in Europe is a subset of that found in the Near East. Elevated diversity values in Iberia may be explained by an limited African influence (CYMBRON *et al.* 1999), but possible also influenced by having received genetic input from a both a littoral and a land-based migration of domestic animals from the Fertile Crescent.

The contour maps of admixture coefficients in **Figures 5.16, 5.17 and 5.18** enable a clear visualisation of the way in which admixture has shaped current cattle populations. As each map represents the modern state of one of the descendants of the three probable domesticates of cattle, they provide an easily comprehensible way to visualise probable past movements of animals, and the people with whom them travelled. The fact that these measures correlate between different datasets is especially significant, as the parental populations are not pre-defined.

The African admixture co-efficient, q_a, clearly highlights the areas of Africa where the original taurine cattle populations remain, predominantly in the south and west of the continent. It is particularly interesting that the combined data enables us to visualise traces influence of African genetic material on southern Iberia, and extending into the Near East. Mitochondrial DNA (RANDO *et al.* 1998; RICHARDS *et al.* 2003) and Y-chromosome (CRUCIANI *et al.* 2002) studies of human populations from Arabia, Europe and Africa have detected significant gene flow between Arabia and East Africa. For example, there is evidence for the presence of North African mtDNA haplotypes in Arab populations of the

Near East. The Y-chromosome influence from Africa to the Near East appears to be more ancient in origin as it is found in both Arab and non-Arab human populations (RICHARDS *et al.* 2003).

This admixture coefficient does not really shed much light on the which of the two models for the initial arrival of cattle in South Africa is more likely: whether they came from the Bantu heartland with Iron Age herders or from the horn of Africa (SMITH 2000). However, it is congruent with the theory that the original cattle in both the west and south were early African domesticates.

The massive influx of *B. indicus* genes from Asia to the Near East, and especially to the east coast of Africa can be clearly seen in **Figure 5.18** (q_i). This introgression has ultimately extended far into the west of the African continent. The signature of genetic exchange from Arabia to Africa is also present in humans. Studies carried out on East African human populations have shown that Ethiopians differ from other Sub-Saharan African human populations as they have elevated frequencies of a mtDNA and Y-chromosome haplogroups which occurs at a high frequency in most Near Eastern populations (RICHARDS *et al.* 2003).

Marshall (2000) notes that the more recent influx of *B. indicus* from Arabia to Africa was the result of an active trade, rather that migrating people. Most of the recent input of human genes from Africa to Arabia seems to have taken place recently and probably traces to women that were brought from Africa as part of the slave trade and then assimilated into Near Eastern populations through inter-marriage and manumission (RICHARDS *et al.* 2003). This fits with a pattern of extensive trade between the two regions. However, it now seems

likely that the major route of *B. indicus* introgression into African was directly to the Horn of Africa and possibly by a maritime route (HANOTTE *et al.* 2002). Interestingly, the signature of minor indicine introductions into Italy may also be evident in **Figure 5.18**. A percentage of individuals of the Modicana breed were found to have bifid processes in the last thoracic vertebrae, traditionally considered a *B. indicus* anatomical characteristic (GRIGSON 2000; PAYNE and HODGES 1997).

The European and Near Eastern coefficient, q_e, displays significant influence in North Africa, where the Moroccan Blonde breed has almost no detectable African component. This may be the result of quite ancient reciprocal trading around the Mediterranean area, which has also lead to the presence of African genetic material in Iberia, as mentioned previously (CYMBRON *et al.* 1999). Obviously each of these admixture maps (**Figures 5.16, 5.17 and 5.18**) is related to the two others, as the increase of one ancestral influence in an area must be associated with the reduction in another type. However, when taken together, they do provide a comprehensive picture of admixture in the total sample region.

5.4.3 METHODOLOGICAL ISSUES.

The accuracy of the method to combine data described here depends both on the predictive power of the data, and the significance of the relationship between the variable in question in the two datasets to be combined. In the case of the datasets used in the current analysis the most critical parameter for successful regression is to have many populations held in common between the two datasets. It is also important that the common populations span a range of values for the variable in question, and especially that low and high values are included. There should be excellent correlation between the values at the low and high end of the scales as it may not be possible to extrapolate outside the range of known values.

Obviously interpolation using a line that is well populated with points is preferable to extrapolation, especially when it is not possible to confirm a linear relationship.

In some cases here the samples that were used for the common breeds were identical samples typed for the different sets of markers, whereas in other cases they were different samples from the same breed. For example, the 35 individuals from the Kuri breed of Chad that were genotyped in dataset 1 are a subset of the 52 individuals that were genotyped in dataset 2. They were selected as the purest examples of the breed and hence the estimates comparing these two are often not highly correlated. The Sahiwal samples in datasets 1 and 2 are derived from different sources. In dataset 1 the animals were sampled in East African and in dataset 2 they are from the Indian sub-continent. Despite these discrepancies, the underlying relationship between the two datasets is strong enough to be detected and the data combining method tolerates this noise.

The limitations associated with this method of data combining are more than outweighed by the benefits of being able to get a global overview of genetic diversity and admixture measures. As admixture is the result of migration and introgression events, the ability to get a global picture of admixture levels and clines of these is particularly interesting and informative. The next step may be to extend this method so that matrices of variables derived from different datasets, including measure like genetic distance, could be combined using common breeds as an anchor to determine the relationship between the matrices. As datasets grow, and estimates of variables become more accurate, this method should also increase in accuracy as the ability of one variable to predict another improves. In terms of the study of genetic diversity in cattle, future plans will involve genotyping the breeds of China, Mongolia and Bhutan. These data may then be combined with previously

published results from Japanese cattle (KIM *et al.* 2002) to produce even more extensive maps of global admixture and diversity levels.

CHAPTER 6

REFERENCES

ABECASIS, G. R., E. NOGUCHI, A. HEINZMANN, J. A. TRAHERNE, S. BHATTACHARYYA *et al.*, 2001 Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet **68:** 191-197.

ALLCHIN, B., and F. R. ALLCHIN, 1982 *The Rise of Civilization in India and Pakistan*. Cambridge University Press, Cambridge.

ALLCHIN, F. R., 1969 Early domestic animals in India and Pakistan, pp. 317-322 in *The Domestication and Exploitation of Plants and Animals*, edited by P. UCKO and G. DIMBLEBY. Duckworth, London.

AMARAL, M. E., S. R. KATA and J. E. WOMACK, 2002 A radiation hybrid map of bovine X chromosome (BTAX). Mamm Genome **13:** 268-271.

AMARAL, M. E. J., S. KATA and J. E. WOMACK, 2000 A radiation hybrid map of BTAX. ISAG 2000 poster B037.

ANDERSON, E. C., 1999 *Lecture Notes for Stat 578C*. Statistical Genetics, Berkeley, University of California.

BALAZS, N. L., J. C. BROWNE, J. D. LOUCK and D. S. STROTTMAN, 2000 Nicholas Constantine Metropolis. Physics Today **53**.

BALLOUX, F., H. BRUNNER, N. LUGON-MOULIN, J. HAUSSER and J. GOUDET, 2000 Microsatellites can be misleading: an empirical and simulation study. Evolution Int J Org Evolution **54:** 1414-1422.

BALLOUX, F., and N. LUGON-MOULIN, 2002 The estimation of population differentiation with microsatellite markers. Mol Ecol **11:** 155-165.

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:** 1359-1368.

BARNHOLTZ-SLOAN, J. S., M. DE ANDRADE and R. CHAKRABORTY, 2001 The impact of population admixture on traditional linkage analysis. Ethn Dis **11:** 519-531.

BARTON, N. H., and K. S. GALE, 1993 Genetic analysis of hybrid zones, pp. 13-45 in *Hybrid zones and the evolutionary process*, edited by R. G. HARRISON. Oxford University Press, Oxford.

BARTON, N. H., and M. SHPAK, 2000 The effect of epistasis on the structure of hybrid zones. Genet Res **75:** 179-198.

BEAUMONT, M., E. M. BARRATT, D. GOTTELLI, A. C. KITCHENER, M. J. DANIELS *et al.*, 2001 Genetic diversity and introgression in the Scottish wildcat. Mol Ecol **10:** 319-336.

BERTORELLE, G., and L. EXCOFFIER, 1998 Inferring admixture proportions from molecular data. Mol Biol Evol **15:** 1298-1311.

BIERNE, N., P. BORSA, C. DAGUIN, D. JOLLIVET, F. VIARD *et al.*, 2003 Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. Mol Ecol **12:** 447-461.

BÖKÖYNI, S., 1989 Definitions of Animal Domestication in *The Walking Larder. Patterns of Domestication, Pastoralism, and Predation*, edited by J. CLUTTON-BROCK. Unwin Hyman Ltd, London.

BÖKÖYNI, S., 1990 Kamid el-Loz. 12. Tierhaltung und Jagd, Saarbrücker Beiträge zur Altertumskunde band 42. Bonn: Dr. Rudolf Habelt GMBH.

BÖKÖYNI, S., 1997 Zebus and Indian Wild Cattle. Anthropozoologica **25-26:** 647-654.

BONNEN, P. E., P. J. WANG, M. KIMMEL, R. CHAKRABORTY and D. L. NELSON, 2002 Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res **12:** 1846-1853.

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:** 455-457.

BRADLEY, D. G., R. T. LOFTUS, P. CUNNIGHAM and D. E. MACHUGH, 1998 The origins of domestic cattle. Evol Anthro.

BRADLEY, D. G., D. E. MACHUGH, P. CUNNINGHAM and R. T. LOFTUS, 1996 Mitochondrial diversity and the origins of African and European cattle. Proc Natl Acad Sci U S A **93:** 5131-5135.

BRADLEY, D. G., D. E. MACHUGH, R. T. LOFTUS, R. S. SOW, C. H. HOSTE *et al.*, 1994 Zebu-taurine variation in Y chromosomal DNA: a sensitive assay for genetic introgression in west African trypanotolerant cattle populations. Anim Genet **25:** 7-12.

BRINKMANN, B., M. KLINTSCHAR, F. NEUHUBER, J. HUHNE and R. BURKHARD, 1998 Mutation rate in Human microsatellites: Influence of the Structure and length of the tandem repeat. Am J Hum Genet **62:** 1408-1415.

BRISCOE, D., J. C. STEPHENS and S. J. O'BRIEN, 1994 Linkage disequilibrium in admixed populations: applications in gene mapping. J Hered **85:** 59-63.

CARDON, L. R., and G. R. ABECASIS, 2003 Using haplotype blocks to map human complex trait loci. Trends Genet **19:** 135-140.

CARLSON, C. S., M. A. EBERLE, M. J. RIEDER, J. D. SMITH, L. KRUGLYAK *et al.*, 2003 Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet **33:** 518-521.

CAVALLI-SFORZA, L. L., 1996 The spread of agriculture and nomadic pastoralism: insights from genetics, linguistics and archaeology in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, edited by D. R. HARRIS. UCL Press Ltd, London.

CAVALLI-SFORZA, L. L., P. MENOZZI and A. PIAZZA, 1993 *The History and Geography of Human Genes*. Princeton University Press.

CHAKRABARTI, D. K., 1990 *The External Trade of the Indus Civilization*. Munshiram Manoharlal, New Delhi.

CHAKRABARTI, D. K., 1997 *Colonial Indology. Sociopolitics of the Ancient Indian Past*. Munshiram Manoharlal, New Delhi.

CHAKRABORTY, A., 2002 Markov Chain Monte Carlo. Resonance **7:** 25-34.

CHAKRABORTY, R., 1975 Estimation of race admixture--a new method. Am J Phys Anthropol **42:** 507-511.

CHAKRABORTY, R., 1986 Gene admixture in human populations: models and predictions. Yearbook of Phys. Anthropology **29:** 1-43.

CHAKRABORTY, R., M. I. KAMBOH and R. E. FERRELL, 1991 'Unique' alleles in admixed populations: a strategy for determining 'hereditary' population differences of disease frequencies. Ethn Dis **1:** 245-256.

CHAKRABORTY, R., M. I. KAMBOH, M. NWANKWO and R. E. FERRELL, 1992 Caucasian genes in American blacks: new data. Am J Hum Genet **50:** 145-155.

CHAKRABORTY, R., and P. E. SMOUSE, 1988 Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. Proc Natl Acad Sci U S A **85:** 3071-3074.

CHAKRABORTY, R., and K. M. WEISS, 1988 Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U S A **85:** 9119-9123.

CHAPMAN, N. H., and E. A. THOMPSON, 2001 Linkage disequilibrium mapping: the role of population history, size, and structure. Adv Genet **42:** 413-437.

CHAPMAN, N. H., and E. A. THOMPSON, 2002 The effect of population history on the lengths of ancestral chromosome segments. Genetics **162:** 449-458.

CHARLES, M. P., 1993 Botanical remains, pp. 203-207 in *Abu Salabikh excavations. The 6G ash-tip and its contents: cultic and administrative discard from the temple*, edited by A. GREEN. British School of Archaeology in Iraq, London.

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

CLASON, A. T., 1978 Late Bronze Age-Iorn Age Zebu in Jordan? Journal of Archaeological Science **5:** 91-93.

CLUTTON-BROCK, J., 1989 Cattle in Ancient North Africa in *The Walking Larder. Patterns of Domestication, Pastoralism, and Predation*, edited by J. CLUTTON-BROCK. Unwin Hyman Ltd, London.

CLUTTON-BROCK, J., 1999 *A Natural History of Domesticated Mammals*. Cambridge University Press, Cambridge.

CLUTTON-BROCK, J., 2000 Cattle, sheep, and goats south of the Sahara: an archaeozoological perspective in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

COLLINS, A., C. LONJOU and N. E. MORTON, 1999 Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci U S A **96:** 15173-15177.

COLLINS-SCHRAMM, H. E., C. M. PHILLIPS, D. J. OPERARIO, J. S. LEE, J. L. WEBER *et al.*, 2002 Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet **70:** 737-750.

CONNAH, G., 2001 *African civilisations: An archaeological perspective*. Cambridge University Press, Cambridge.

CRUCIANI, F., P. SANTOLAMAZZA, P. SHEN, V. MACAULAY, P. MORAL *et al.*, 2002 A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet **70:** 1197-1214.

CSEP, 1996 *Computational Science Education Project*. http://csep1.phy.ornl.gov/.

CYMBRON, T., 2003 *Molecular Genetic Analysis of Mediterranean Domestic Cattle breeds*. Ph.D. Thesis, University of Lisbon.

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.

DEAN, M., J. C. STEPHENS, C. WINKLER, D. A. LOMB, M. RAMSBURG *et al.*, 1994 Polymorphic admixture typing in human ethnic populations. Am J Hum Genet **55:** 788-808.

DENG, H., W. CHEN and R. RECKER, 2001 Population Admixture. Detection by hardy-weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. Genetics **157:** 885-897.

DESAI, P. Y., and P. W. GLYNN, 2001 A Markov Chain perspective on Adaptive Monte Carlo algorithms, pp. 379-384 in *Proceeding of the Winter Simulation Conference*, edited by J. S. S. B. A. PETERS, D. J. MEDEIROS, AND M. W. ROHRER, Arlington, VA, USA.

DESTRO-BISOL, G., R. MAVIGLIA, A. CAGLIA, I. BOSCHI, G. SPEDINI *et al.*, 1999 Estimating European admixture in African Americans by using microsatellites and a microsatellite haplotype (CD4/Alu). Hum Genet **104:** 149-157.

DI BENEDETTO, G., A. ERGUVEN, M. STENICO, L. CASTRI, G. BERTORELLE *et al.*, 2001 DNA diversity and population admixture in Anatolia. Am J Phys Anthropol **115:** 144-156.

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:** 3166-3170.

DIAMOND, J., 1997 *Guns, Germs and Steel.* Norton, New York.

DROEGEMUELLER, C., O. DISTL and T. LEEB, 2000 Identification of a highly polymorphic microsatellite within the bovine ectodysplasin A (ED1) gene on BTA Xq22-24. Anim Genet **31:** 416-416.

DUNNING, A. M., F. DUROCHER, C. S. HEALEY, M. D. TEARE, S. E. MCBRIDE *et al.*, 2000 The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet **67:** 1544-1554.

DUPANLOUP, I., and G. BERTORELLE, 2001 Inferring admixture proportions from molecular data: extension to any number of parental populations. Mol Biol Evol **18:** 672-675.

EDWARDS, A., H. A. HAMMOND, L. JIN, C. T. CASKEY and R. CHAKRABORTY, 1992 Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics **12:** 241-253.

EDWARDS, C. J., 2003 *Genetic Variation in Domesticated Cattle and Wild Aurochsen using Ancient and Modern DNA.* Ph. D. thesis, Trinity College, University of Dublin.

EDWARDS, C. J., G. DOLF, C. LOOFT, R. T. LOFTUS and D. G. BRADLEY, 2000 Relationships between the endangered Pustertaler-Sprinzen and three related European cattle breeds as analysed with 20 microsatellite loci. Anim Genet **31:** 329-332.

EISLER, M. C., J. BRANDT, BAUER B, P. H. CLAUSEN, V. DELESPAUX *et al.*, 2001 Standardised tests in mice and cattle for the detection of drug resistance in tsetse-transmitted trypanosomes of African domestic cattle. Vet Parasitol **97:** 171-182.

ELLEGREN, H., 1994 Linkage mapping of the apolipoprotein A-I (APOA1) gene to pig chromosome 9. Mamm Genome **5:** 58-59.

EPSTEIN, H., 1971 *The Origin of the Domestic Animals of Africa.* Africana, New York.

EPSTEIN, H., and I. L. MASON, 1984 Cattle, pp. 6-28 in *Evolution of Domestic Animals*, edited by I. L. MASON. Longman, London and New York.

ESTOUP, A., J. M. CORNUET, F. ROUSSET and R. GUYOMARD, 1999 Juxtaposed microsatellite systems as diagnostic markers for admixture: theorical aspects. Molecular Biology and Evolution **16:** 898-908.

ESTOUP, A., L. GARNERY, M. SOLIGNAC and J. M. CORNUET, 1995a Microsatellite variation in honey bee (Apis mellifera L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics **140:** 679-695.

ESTOUP, A., C. TAILLIEZ, J. M. CORNUET and M. SOLIGNAC, 1995b Size homoplasy and mutational processes of interrupted microsatellites in two bee species, Apis mellifera and Bombus terrestris (Apidae). Mol Biol Evol **12:** 1074-1084.

FALUSH, D., M. STEPHENS and J. K. PRITCHARD, 2003 Inference of population structure II: Linked loci and correlated allele frequencies. (in prep).

FAO, 2003a Global Strategy for the Management of Farm Animal Genetic Resources. http://dad.fao.org.

FAO, 2003b The Program against African Trypanosomiasis, http://www.fao.org/paat/html/home.htm.

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

FARNIR, F., B. GRISART, W. COPPIETERS, J. RIQUET, P. BERZI *et al.*, 2002 Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics **161:** 275-287.

FELIUS, M., 1995 *Cattle breeds - an Encyclopedia*. Misset, Doetinchem, Netherlands.

FELSENSTEIN, J., 2001 PHYLIP (the PHYLogeny Inference Package), University ofWashington, USA.

FREIMER, N. B., S. K. SERVICE and M. SLATKIN, 1997 Expanding on population studies. Nat Genet **17:** 371-373.

FULLER, D. Q., 2001 Fifty Years of Archaeobotanical Studies in India: Laying a Solid Foundation, pp. 247-363 in *Indian Archaeology in Retrospect, Archaeology and Interactive Disciplines*, edited by S. SETTAR and R. KORISETTAR. Publications of the Indian Council for Historical Research, Manohar, New Dehli.

GABRIEL, S. B., S. F. SCHAFFNER, H. NGUYEN, J. M. MOORE, J. ROY *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225-2229.

GARNER, C., and M. SLATKIN, 2003 On selecting markers for association studies: Patterns of linkage disequilibrium between two and three diallelic loci. Genet Epidemiol **24:** 57-67.

GARZA, J. C., M. SLATKIN and N. B. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol **12:** 594-603.

GIFFORD-GONZALEZ, D., 2000 Animal Disease Challenges to the Emergence of Pastoralism in Sub-Saharan Africa. African Archaeological Review **17:** 95-139.

GOLDSTEIN, D. B., and A. G. CLARK, 1995 Microsatellite variation in North American populations of Drosophila melanogaster. Nucleic Acids Res **23:** 3882-3886.

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:** 463-471.

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:** 6723-6727.

GOLDSTEIN, D. B., and C. SCHLOTTERER, 1999 *Microsatellites: Evolution and Applications*. Oxford University Press, London.

GRIGSON, C., 2000 *Bos africanus* (Brehm)? Notes on the archaeozoology of the native cattle of Africa in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

HANOTTE, O., D. G. BRADLEY, J. W. OCHIENG, Y. VERJEE, E. W. HILL *et al.*, 2002 African pastoralism: genetic imprints of origins and migrations. Science **296:** 336-339.

HANOTTE, O., C. L. TAWAH, D. G. BRADLEY, M. OKOMO, Y. VERJEE *et al.*, 2000 Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-saharan African cattle breeds. Mol Ecol **9:** 387-396.

HANSEN, M. M., E. E. NEILSEN, D. BEKKEVOLD and K. L. MENSBERG, 2001 Admixture analysis and stocking impact assessment in brown trout (Salmo trutta), estimated with incomplete baseline data. Can J Fish Aquat Sci **58:** 1853-1860.

HARTL, D. L., and A. G. CLARK, 1997 *Principles of Population Genetics*. Sinauer, Sunderland, MA.

HASSAN, F. A., 2000 Climate and cattle in North Africa: a first approximation in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

HEDRICK, P. W., 1999 Highly Variable Loci and their interpretation in Evolution and Conservation. Evolution **53:** 313-318.

HEDRICK, P. W., 2000 *Genetics of Populations*. Jones and Bartlett, London.

HIGGINS, D. G., and P. M. SHARP, 1988 CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73:** 237-244.

HORTON, 1987 The Swahili Corridor. Sci Amer **257:** 76-84.

HUTTLEY, G. A., M. W. SMITH, M. CARRINGTON and S. J. O'BRIEN, 1999 A scan for linkage disequilibrium across the human genome. Genetics **152:** 1711-1722.

HUTTLEY, G. A., and S. R. WILSON, 2000 Testing for concordant equilibria between population samples. Genetics **156:** 2127-2135.

ISNAR, 1995 *Survival in the Sahel: An ecological and developmental challenge*. International Service for National Agricultural Research.

JIN, L., and R. CHAKRABORTY, 1994 Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. Mol Biol Evol **11:** 120-127.

JORDE, L. B., M. BAMSHAD and A. R. ROGERS, 1998 Using mitochondrial and nuclear DNA markers to reconstruct human evolution. Bioessays **20:** 126-136.

JORDE, L. B., A. R. ROGERS, M. BAMSHAD, W. S. WATKINS, P. KRAKOWIAK *et al.*, 1997 Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci U S A **94:** 3100-3103.

KABAYO, J. P., 2002 Aiming to eliminate tsetse from Africa. Trends Parasitol **18:** 473-475.

KALINOWSKI, S. T., and P. W. HEDRICK, 2001 Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep. Heredity **87:** 698-708.

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:** 446-457.

KAPLAN, N. L., E. R. MARTIN, R. W. MORRIS and B. S. WEIR, 1998 Marker selection for the transmission/disequilibrium test, in recently admixed populations. Am J Hum Genet **62:** 703-712.

KEMP, S., 2002 The Wellcome Funded Host Pathogen Interaction Program. http://www.genomics.liv.ac.uk/tryps/.

KIM, K. S., J. S. YEO and C. B. CHOI, 2002 Genetic diversity of north-east Asian cattle based on microsatellite data. Anim Genet **33:** 201-204.

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

KIMURA, M., and T. OHTA, 1978 Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci U S A **75:** 2868-2872.

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:** 1142-1145.

KRUGLYAK, L., 1999a Genetic isolates: separate but equal? Proc Natl Acad Sci U S A **96:** 1170-1172.

KRUGLYAK, L., 1999b Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet **22:** 139-144.

KUMAR, P., A. FREEMAN, R. LOFTUS, G. C, D. FULLER *et al.*, 2003 Admixture analysis of South Asian Cattle. Heredity **in press.**

LAAN, M., and S. PAABO, 1997 Demographic history and linkage disequilibrium in human populations. Nat Genet **17:** 435-438.

LAAN, M., and S. PAABO, 1998 Mapping genes by drift-generated linkage disequilibrium. Am J Hum Genet **63:** 654-656.

LAGZIEL, A., E. LIPKIN and M. SOLLER, 1996 Association between SSCP haplotypes at the bovine growth hormone gene and milk protein percentage. Genetics **142:** 945-951.

LAMBERG-KARLOVSKY, C. C., 1989 Mesopotamia, Central Asia and the Indus Valley: so the kings were killed, pp. 241-267 in *Archaeological Thought in America*, edited by C. C. LAMBERG-KARLOVSKY. Cambridge University Press, Cambridge.

LAUTENBERGER, J. A., J. C. STEPHENS, S. J. O'BRIEN and M. W. SMITH, 2000 Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet **66:** 969-978.

LEGROS, D., G. OLLIVIER, M. GASTELLU-ETCHEGORRY, C. PAQUET, C. BURRI *et al.*, 2002 Treatment of human African trypanosomiasis--present situation and needs for research and development. Lancet Infect Dis **2:** 393.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol **4:** 203-221.

LEVTZION, N., 1973 *Ancient Ghana and Mali*. Methuen, London.

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:** 2015-2022.

LOFTUS, R. T., D. E. MACHUGH, D. G. BRADLEY, P. M. SHARP and P. CUNNINGHAM, 1994a Evidence for two independent domestications of cattle. Proc Natl Acad Sci U S A **91:** 2757-2761.

LOFTUS, R. T., D. E. MACHUGH, L. O. NGERE, D. S. BALAIN, A. M. BADI *et al.*, 1994b Mitochondrial genetic variation in European, African and Indian cattle populations. Anim Genet **25:** 265-271.

LONG, J. C., 1991 The genetic structure of admixed populations. Genetics **127:** 417-428.

LONJOU, C., A. COLLINS and N. E. MORTON, 1999 Allelic association between marker loci. Proc Natl Acad Sci U S A **96:** 1621-1626.

MACDONALD, K. C., 1998 Before the Empire of Ghana: pastoralism and the origins of cultural complexity in the Sahel., pp. 71-103 in *Transformations in Africa: essays on Africa's later past*, edited by G. CONNAH. Leicester University Press, London and Washington.

MACDONALD, K. C., 2000 The origins of African livestock: indigenous or imported? in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

MACDONALD, K. C., and R. HUTTON MACDONALD, 2000 The origins and development of domesticated animals in arid West Africa in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

MACHUGH, D. E., 1996 *Molecular Biogeography and Genetic Structure of Domesticated Cattle*. Ph.D. Thesis, Trinity College, University of Dublin.

MACHUGH, D. E., C. J. EDWARDS, J. F. BAILEY, D. R. BANCROFT and D. G. BRADLEY, 2000 The extraction and analysis of ancient DNA from bone and teeth: a survey of current methodologies. Ancient Biomolecules **3:** 81-102.

MACHUGH, D. E., R. T. LOFTUS, D. G. BRADLEY, P. M. SHARP and P. CUNNINGHAM, 1994 Microsatellite DNA variation within and among European cattle breeds. Proc R Soc Lond B Biol Sci **256:** 25-31.

MACHUGH, D. E., R. T. LOFTUS, P. CUNNINGHAM and D. G. BRADLEY, 1998 Genetic structure of seven European cattle breeds assessed using 20 microsatellite markers. Anim Genet **29:** 333-340.

MACHUGH, D. E., M. D. SHRIVER, R. T. LOFTUS, P. CUNNINGHAM and D. G. BRADLEY, 1997 Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*). Genetics **146:** 1071-1086.

MAGEE, D. A., 2002 *Molecular Genetic Investigations of the Diversity and Origins of Old and New World Cattle*. Ph. D. Thesis, Trinity College, University of Dublin.

MALASPINA, P., F. CRUCIANI, P. SANTOLAMAZZA, A. TORRONI, A. PANGRAZIO *et al.*, 2000 Patterns of male-specific inter-population divergence in Europe, West Asia and North Africa. Ann Hum Genet **64:** 395-412.

MANLY, B. F. G., 1986 *Multivariate Statistical Methods: A Primer*.

MANNEN, H., S. TSUJI, R. T. LOFTUS and D. G. BRADLEY, 1998 Mitochondrial DNA variation and evolution of Japanese black cattle (*Bos taurus*). Genetics **150:** 1169-1175.

MARSHALL, F., 2000 The origins and spread of domestic animals in East Africa in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

MARSHALL, F., and E. HILDEBRAND, 2002 Cattle before crops:The beginnings of Food production in Africa. Journal of World Prehistory **16:** 99-143.

MARTH, G. T., I. KORF, M. D. YANDELL, R. T. YEH, Z. GU *et al.*, 1999 A general approach to single-nucleotide polymorphism discovery. Nat Genet **23:** 452-456.

MATOVU, E., T. SEEBECK, J. C. ENYARU and R. KAMINSKY, 2001 Drug resistance in Trypanosoma brucei spp., the causative agents of sleeping sickness in man and nagana in cattle. Microbes Infect **3:** 763-770.

MCINTOSH, R. J., 1993 The Pulse Model: Genesis and Accommodation of Specialization in Middle Niger. Journal of African History **34:** 181-220.

MCINTOSH, S. K., and R. J. MCINTOSH, 1979 Initial perspectives on prehistoric subsistence in the Inland Niger Delta (Mali). World Archaeology **11:** 227-243.

MCINTOSH, S. K., and R. J. MCINTOSH, 1980 *Prehistoric investigations in the region of Jenné, Mali*. Cambridge Monographs in African Archaeology, BAR International Series, Oxford.

MCKEIGUE, P. M., 1997 Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet **60:** 188-196.

MCKEIGUE, P. M., 1998 Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet **63:** 241-251.

MCKEIGUE, P. M., J. R. CARPENTER, E. J. PARRA and M. D. SHRIVER, 2000 Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann Hum Genet **64:** 171-186.

MCPEEK, M. S., and A. STRAHS, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet **65:** 858-875.

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

MEADOW, R. H., 1987 Faunal Exploitation Patterns in Eastern Iran and Baluchistan: A Review of Recent Investigations, pp. 881-916 in *Oreintalia Iosephi Tucci Memoriae Dicata. Serie Orientale Roma LVI, 2, Istituto Italiano per il Medio ed Estermo Oreinte*, edited by G. GNOLI and L. LANCIOTTI, Rome.

MEADOW, R. H., 1996 The origins and spread of agriculture and pastoralism in northwestern South Asia, pp. 390-412 in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, edited by D. R. HARRIS. UCL Press Ltd, London.

MEDJUGORAC, I., W. KUSTERMANN, P. LAZAR, I. RUSS and F. PIRCHNER, 1994 Marker-derived phylogeny of European cattle supports demic expansion of agriculture. Anim Genet **25 Suppl 1:** 19-27.

MEGHEN, C., D. E. MACHUGH and D. G. BRADLEY, 1994 *Genetic Characterisation and West African cattle.* FAO, Rome.

MINCH, E., A. RUIZ-LINARES, D. GOLDSTEIN, B. , M. W. FELDMAN and L. L. CAVALLI-SFORZA, 1995 MICROSAT: The Microsatellite Distance program.

MOAZAMI-GOUDARZI, K., and D. LALOE, 2002 Is a multivariate consensus representation of genetic relationships among populations always meaningful? Genetics **162:** 473-484.

MOUNTAIN, J. L., A. A. LIN, A. M. BOWCOCK and L. L. CAVALLI-SFORZA, 1992 Evolution of modern humans: evidence from nuclear DNA polymorphisms. Philos Trans R Soc Lond B Biol Sci **337:** 159-165.

NEEL, J. V., 1973 "Private" genetic variants and the frequency of mutation among South American Indians. Proc Natl Acad Sci U S A **70:** 3311-3315.

NEI, M., 1972 Genetic distance between populations. Amer Nat **106:** 283-292.

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

NORDBORG, M., and S. TAVARE, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet **18:** 83-90.

ORTI, G., D. E. PEARSE and J. C. AVISE, 1997 Phylogenetic assessment of length variation at a microsatellite locus. Proc Natl Acad Sci U S A **94:** 10745-10749.

OTT, J., 2000 Predicting the range of linkage disequilibrium. Proc Natl Acad Sci U S A **97:** 2-3.

PAABO, S., 2003 The mosaic that is our genome. Nature **421:** 409-412.

PAETKAU, D., L. P. WAITS, P. L. CLARKSON, L. CRAIGHEAD and C. STROBECK, 1997 An empirical evaluation of genetic distance statistics using microsatellite data from bear (*Ursidae*) populations. Genetics **147:** 1943-1957.

PARK, S. D., 2001 *Trypanotolerance in West African cattle and the Population Genetic Effects of Selection.* Ph.D. Thesis, Trinity College, University of Dublin.

PARPOLA, A., 1994 *Deciphering the Indus script.* Cambridge University Press, Cambridge.

PARRA, E. J., R. A. KITTLES, G. ARGYROPOULOS, C. L. PFAFF, K. HIESTER *et al.*, 2001 Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol **114:** 18-29.

PARRA, E. J., A. MARCINI, J. AKEY, J. MARTINSON, M. A. BATZER *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet **63:** 1839-1851.

PAYNE, W. J. A., 1970 *Cattle Production in the Tropics.* Longman, London.

PAYNE, W. J. A., and J. HODGES, 1997 *Tropical Cattle:Origins, Breeds and Breeding Policies.* Blackwell Science Ltd., Oxford.

PAYNE, W. J. A., and R. T. WILSON, 1999 *An introduction to Animal Husbandry in the Tropics.* Blackwell Science Ltd., Oxford.

PETERSON, A. C., A. DI RIENZO, A. E. LEHESJOKI, A. DE LA CHAPELLE, M. SLATKIN *et al.*, 1995 The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet **4:** 887-894.

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

PRITCHARD, J. K., and M. PRZEWORSKI, 2001 Linkage disequilibrium in humans: models and data. Am J Hum Genet **69:** 1-14.

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

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

RANDO, J. C., F. PINTO, A. M. GONZALEZ, M. HERNANDEZ, J. M. LARRUGA *et al.*, 1998 Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. Ann Hum Genet **62 (Pt 6):** 531-550.

RANNALA, B., 2001 Finding genes influencing susceptibility to complex diseases in the post-genome era. Am J Pharmacogenomics **1:** 203-221.

RANNALA, B., and M. SLATKIN, 2000 Methods for multipoint disease mapping using linkage disequilibrium. Genet Epidemiol **19 Suppl 1:** S71-77.

REED, T. E., 1973 Number of gene loci required for accurate estimation of ancestral population proportions in individual human hybrids. Nature **244:** 575-576.

RENFREW, C., 1987 *Archaeology and Language - The Puzzle of Indo-European Origins.* Cambridge University Press, Cambridge.

RENFREW, C., 1996 Language families and the spread of farming in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, edited by D. R. HARRIS. UCL Press Ltd, London.

RICHARDS, M., C. RENGO, F. CRUCIANI, F. GRATRIX, J. F. WILSON *et al.*, 2003 Extensive female-mediated gene flow from sub-saharan Africa into near eastern arab populations. Am J Hum Genet **72:** 1058-1064.

RITZ, L. R., M. L. GLOWATZKI-MULLIS, D. E. MACHUGH and C. GAILLARD, 2000 Phylogenetic analysis of the tribe Bovini using microsatellites. Anim Genet **31:** 178-185.

ROBERTS, D. F., and R. W. HIORNS, 1965 Methods of analysis of the genetic composition of a hybrid population. Hum Biol **37:** 38-43.

ROBERTSHAW, P., 1989 The development of pastoralism in East Africa in *The Walking Larder. Patterns of Domestication, Pastoralism, and Predation*, edited by J. CLUTTON-BROCK. Unwin Hyman Ltd, London.

ROGERS, A. R., 1995 Genetic evidence for a Pleistocene population explosion. Evolution **49:** 608-615.

ROGERS, A. R., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol **9:** 552-569.

ROGERS, D. J., and S. E. RANDOLPH, 2002 A response to the aim of eradicating tsetse from Africa. Trends Parasitol **18:** 534-536.

ROSENBERG, N. A., T. BURKE, K. ELO, M. W. FELDMAN, P. J. FREIDLIN *et al.*, 2001a Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics **159:** 699-713.

ROSENBERG, N. A., E. WOOLF, J. K. PRITCHARD, T. SCHAAP, D. GEFEL *et al.*, 2001b Distinctive genetic signatures in the Libyan Jews. Proc Natl Acad Sci U S A **98:** 858-863.

ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD *et al.*, 2002 Genetic structure of human populations. Science **298:** 2381-2385.

RUBINSZTEIN, D. C., W. AMOS, J. LEGGO, S. GOODBURN, S. JAIN *et al.*, 1995 Microsatellite evolution--evidence for directionality and variation in rate between species. Nat Genet **10:** 337-343.

RYBICKI, B. A., S. K. IYENGAR, T. HARRIS, R. LIPTAK, R. C. ELSTON *et al.*, 2002 The distribution of long range admixture linkage disequilibrium in an african-american population. Hum Hered **53:** 187-196.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:** 406-425.

SCHLOTTERER, C., 2000 Evolutionary dynamics of microsatellite DNA. Chromosoma **109:** 365-371.

SCHLOTTERER, C., and D. TAUTZ, 1992 Slippage synthesis of simple sequence DNA. Nucleic Acids Res **20:** 211-215.

SCHNEIDER, S., D. ROESSLI and L. EXCOFFIER, 2000 Arlequin ver. 2.000: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

SHAFFER, J. G., 1984 The Indo-Aryan invasions: cultural myth and archaeological reality, pp. 77-90 in *The People of South Asia: The Biological Anthropology of India, Pakistan and Nepal*, edited by J. R. LUKACS. Plenum, London.

SHERRINGTON, R., G. MELMER, M. DIXON, D. CURTIS, B. MANKOO *et al.*, 1991 Linkage disequilibrium between two highly polymorphic microsatellites. Am J Hum Genet **49:** 966-971.

SHIFMAN, S., J. KUYPERS, M. KOKORIS, B. YAKIR and A. DARVASI, 2003 Linkage disequilibrium patterns of the human genome across populations. Hum Mol Genet **12:** 771-776.

SHRIVER, M. D., L. JIN, E. BOERWINKLE, R. DEKA, R. E. FERRELL *et al.*, 1995 A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol Biol Evol **12:** 914-920.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY and E. BOERWINKLE, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983-993.

SHRIVER, M. D., E. J. PARRA, S. DIOS, C. BONILLA, H. NORTON *et al.*, 2003 Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet **112:** 387–399.

SHRIVER, M. D., M. W. SMITH, L. JIN, A. MARCINI, J. M. AKEY *et al.*, 1997 Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet **60:** 957-964.

SINCLAIR, P. J. J., 1995 The origin of urbanism in East and southern Africa:a diachronic perspective, pp. 99-109 in *Islamic art and culture in sub-Saharan Africa*, edited by K. ÅDAHL and B. SAHLSTRÖM. Acta Universitatis Upsaliensis, Upsalla.

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

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

SMITH, A. B., 2000 The origins of the domesticated animals of southern Africa in *The origins and development of African livestock: archaeology, genetics, linguistics and ethnography*, edited by R. M. BLENCH and K. C. MACDONALD. UCL Press, London.

SONSTEGARD, T. S., W. BARENDSE, G. L. BENNETT, G. A. BROCKMANN, S. DAVIS *et al.*, 2001 Consensus and comprehensive linkage maps of the bovine sex chromosomes. Anim Genet **32:** 115-117.

STICH, A., P. M. ABEL and S. KRISHNA, 2002 Human African trypanosomiasis. Bmj **325:** 203-206.

STUMPF, M. P., 2002 Haplotype diversity and the block structure of linkage disequilibrium. Trends Genet **18:** 226-228.

STUMPF, M. P., and D. B. GOLDSTEIN, 2003 Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. Curr Biol **13:** 1-8.

SUNNUCKS, P., A. C. WILSON, L. B. BEHEREGARAY, K. ZENGER, J. FRENCH *et al.*, 2000 SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. Mol Ecol **9:** 1699-1710.

TAILLON-MILLER, P., I. BAUER-SARDINA, N. L. SACCONE, J. PUTZEL, T. LAITINEN *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat Genet **25:** 324-328.

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

TENESA, A., S. A. KNOTT, D. WARD, D. SMITH, J. L. WILLIAMS *et al.*, 2003 Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. J Anim Sci **81:** 617-623.

TERES, V., D. H. LESTER and J. L. WILLIAMS, 1996 Mapping of bovine fibroblast growth factor receptor 3 (FGFR3) to the telomeric region of chromosome 6 by SSCP analysis. Anim Genet **27:** 371.

THOMAS, J., 1996 The cultural context of the first use of domesticates in continental Central and Northwest Europe, pp. 310-322 in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, edited by D. R. HARRIS. UCL Press Ltd, London.

THOMPSON, E. A., and J. V. NEEL, 1997 Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am J Hum Genet **60:** 197-204.

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:** 4673-4680.

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:** 1088-1091.

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

VARILO, T., T. PAUNIO, A. PARKER, M. PEROLA, J. MEYER *et al.*, 2003 The interval of linkage disequilibrium (LD) detected with microsatellite and SNP markers in chromosomes of Finnish populations with different histories. Hum Mol Genet **12:** 51-59.

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

WEIR, B. S., and W. G. HILL, 2002 Estimating *F*-statistics. Annu Rev Genet **36:** 721-750.

WENDORF, F., and R. SCHILD, 1994 Are the early Holocene cattle in the Eastern Sahara domestic or wild? Evol Anthropol **3**.

WHO, 2002 Human African trypanosomiasis/ Trypanosomiase humaine Africaine., pp. 3 in *WHO/CDS/CSR/EPH/2002.14*.

WIEHE, T., J. MOUNTAIN, P. PARHAM and M. SLATKIN, 2000 Distinguishing recombination and intragenic gene conversion by linkage disequilibrium patterns. Genet Res **75:** 61-73.

WIEHE, T., and M. SLATKIN, 1998 Epistatic selection in a multi-locus Levene model and implications for linkage disequilibrium. Theor Popul Biol **53:** 75-84.

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

WILSON, J. F., and D. B. GOLDSTEIN, 2000 Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. Am J Hum Genet **67:** 926-935.

WRIGHT, S., 1951 The genetical structure of populations. Ann Eugen **15:** 323-254.

WRIGHT, S., 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution **19**.

WRIGHT, S., 1969 *Theory of gene frequencies*. University of Chicago Press,, Chicago, London.

WU, R., and Z. ZENG, 2001 Joint linkage and linkage disequilibrium mapping in natural populations. Genetics **157:** 899-909.

ZAPATA, C., 2000 The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution Int J Org Evolution **54:** 1809-1812.

ZEUNER, F. E., 1963 The history of the domestication of cattle, pp. 9-20 in *Man And Cattle,*, edited by A. E. MOURANT and F. E. ZEUNER. Royal Anthropological Institute of Great Britain and Ireland.

ZILHAO, J., 2001 Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. Proc Natl Acad Sci U S A **98:** 14180-14185.

APPENDIX A

MICROSATELLITE ALLELE FREQUENCIES

Appendix A.1 *Frequencies of ten microsatellite loci in 7 Indian cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

Breed	BM 1818		BM 1824		BM 2113		CSRM 60		CSSM66		ETH 3		ETH 10		ETH 225		Hel 1		Hel 5	
ONGOLE	*256*	3	*179*	18	*130*	26	*93*	21	*179*	22	*103*	10	*207*	21	*141*	1	*101*	19	*151*	26
	258	2	*181*	21	*136*	1	*95*	2	*181*	26	*115*	33	*209*	37	*147*	1	*103*	11	*155*	35
	260	1	*183*	15	*138*	5	*103*	8	*185*	1	*117*	15	*211*	5	*153*	19	*105*	1	*163*	1
	262	17	*189*	1	*140*	2	*105*	13	*187*	2	*125*	2	*217*	1	*157*	38	*107*	28	*165*	2
	264	30	*193*	9	*142*	28	*111*	11	*199*	12					*159*	5	*111*	3		
	266	2			*144*	2	*115*	9	*201*	1							*117*	2		
	270	3																		
	272	5																		
	280	1																		
NELLORE	*258*	1	*179*	24	*130*	25	*93*	22	*179*	19	*103*	10	*207*	17	*143*	1	*101*	9	*151*	28
	262	21	*181*	11	*134*	2	*95*	1	*181*	16	*115*	18	*209*	33	*153*	11	*103*	1	*155*	22
	264	23	*183*	14	*136*	2	*97*	1	*183*	1	*117*	24	*211*	2	*157*	42	*105*	7	*165*	4
	266	3	*195*	5	*138*	1	*99*	3	*185*	15	*125*	2	*217*	2			*107*	18		
	270	5			*140*	2	*103*	25	*199*	3							*111*	8		
	278	1			*142*	12	*113*	1									*117*	11		
					144	10	*115*	1												
RED SINDHI	*256*	1	*179*	2	*130*	10	*91*	2	*179*	30	*103*	10	*207*	34	*139*	4	*101*	7	*151*	37
	262	32	*181*	46	*136*	19	*93*	19	*181*	12	*115*	35	*209*	7	*141*	1	*105*	24	*153*	2
	264	8	*183*	25	*138*	2	*95*	9	*183*	13	*117*	17	*211*	23	*143*	1	*107*	26	*155*	8
	266	6	*195*	1	*140*	18	*97*	1	*185*	1	*119*	1	*213*	1	*149*	5	*111*	11	*165*	5
	268	6			*142*	11	*101*	4	*187*	3	*121*	1	*215*	3	*153*	17	*113*	1	*167*	22
	270	12			*144*	5	*103*	5	*189*	1	*125*	8	*217*	4	*157*	46	*117*	3		
	272	6			*146*	9	*105*	3	*193*	2										
	278	5					*109*	18	*199*	13										
							115	11	*201*	1										

Appendix A.1 (cont) *Frequencies of ten microsatellite loci in 7 Indian cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

Breed	BM 1818		BM 1824		BM 2113		CSRM 60		CSSM66		ETH 3		ETH 10		ETH 225		Hel 1		Hel 5	
THARPAKER	*260*	1	*179*	1	*130*	24	*93*	42	*179*	29	*103*	10	*207*	18	*139*	12	*101*	7	*151*	35
	262	5	*181*	48	*132*	2	*97*	2	*181*	7	*115*	24	*209*	22	*141*	1	*103*	2	*155*	2
	264	48	*183*	15	*136*	14	*103*	3	*183*	2	*117*	23	*211*	25	*149*	1	*107*	14	*165*	25
	266	10	*185*	1	*138*	1	*109*	12	*185*	12	*119*	7	*217*	1	*151*	3	*109*	1		
	270	2	*189*	1	*140*	16	*111*	5	*195*	1					*153*	3	*111*	33		
	272	4	*193*	2	*142*	11	*113*	6	*197*	11					*157*	50	*117*	9		
									199	3										
									209	3										
DESI	*256*	3	*179*	6	*130*	7	*91*	2	*179*	18	*103*	2	*207*	21	*139*	6	*101*	8	*151*	21
	258	2	*181*	20	*136*	9	*93*	15	*181*	12	*115*	22	*209*	16	*143*	2	*105*	9	*155*	23
	260	1	*183*	19	*138*	5	*95*	5	*185*	5	*117*	10	*211*	11	*147*	4	*107*	27	*157*	2
	262	23	*193*	6	*140*	14	*97*	1	*193*	1	*119*	3	*213*	1	*149*	1	*111*	5	*163*	1
	264	9	*195*	1	*142*	16	*103*	7	*195*	1	*125*	3	*215*	1	*153*	8	*113*	1	*167*	5
	266	1			*144*	1	*105*	2	*197*	6			*217*	2	*157*	31				
	268	3					*109*	3	*199*	7										
	270	5					*111*	12	*203*	2										
	272	1					*113*	4												
	280	4					*115*	1												

Appendix A.1 (cont) *Frequencies of ten microsatellite loci in 7 Indian cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

Breed	BM 1818		BM 1824		BM 2113		CSRM 60		CSSM66		ETH 3		ETH 10		ETH 225		Hel 1		Hel 5	
HARIANA	*258*	1	*179*	1	*130*	6	*93*	8	*179*	5	*103*	2	*207*	5	*135*	3	*101*	1	*151*	7
	262	9	*181*	7	*136*	5	*95*	4	*181*	4	*115*	9	*209*	2	*143*	1	*105*	3	*155*	9
	264	5	*183*	10	*140*	3	*103*	2	*183*	1	*117*	7	*211*	9	*147*	4	*107*	13	*163*	1
	266	1	*193*	2	*142*	5	*111*	5	*187*	1	*119*	2	*217*	4	*157*	10	*109*	1	*165*	3
	270	3			*144*	1	*115*	1	*195*	2					*159*	2	*111*	1		
	272	1							*197*	3							*117*	1		
									199	3										
									207	1										
SAHIWAL	*262*	13	*181*	15	*128*	2	*93*	10	*179*	3	*115*	20	*207*	15	*135*	2	*101*	1	*151*	6
	264	1	*183*	1	*130*	3	*95*	1	*181*	4			*209*	1	*139*	3	*105*	1	*155*	2
	266	1	*193*	4	*136*	7	*97*	1	*185*	4			*211*	2	*143*	1	*107*	14	*165*	1
	268	2			*140*	4	*101*	4	*197*	5			*217*	2	*153*	8	*117*	2	*167*	11
	270	3			*142*	4	*105*	1	*199*	4					*157*	6				
							109	1												
							111	1												
							113	1												

Appendix A.2 *Frequencies of nine microsatellite loci in 7 Indian cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

Breed	Hel 9		Hel 13		ILSTS 005		ILSTS 006		INRA 005		INRA 023		INRA 032		INRA 035		INRA 063	
ONGOLE	*149*	17	*182*	39	*182*	21	*289*	1	*137*	27	*195*	7	*156*	2	*102*	10	*175*	4
	151	1	*186*	25	*184*	1	*291*	2	*139*	8	*197*	2	*160*	18	*104*	6	*177*	18
	155	3			*186*	4	*293*	24	*141*	7	*203*	13	*176*	11	*106*	1	*183*	35
	157	21			*190*	27	*295*	17	*143*	19	*209*	1	*178*	3	*108*	11	*185*	5
	161	9			*194*	11	*297*	15	*147*	1	*211*	1	*184*	16	*110*	24		
	163	5					*301*	1	*149*	2	*215*	40	*190*	7	*120*	9		
	165	8											*202*	7	*124*	1		
NELLORE	*149*	3	*182*	32	*182*	18	*291*	1	*137*	29	*195*	4	*160*	14	*102*	6	*177*	8
	151	5	*186*	22	*184*	5	*293*	16	*139*	10	*197*	3	*174*	15	*104*	14	*183*	40
	155	10			*186*	4	*295*	22	*141*	7	*203*	16	*176*	3	*106*	1	*185*	6
	157	19			*190*	20	*297*	12	*143*	5	*209*	8	*178*	10	*110*	21		
	159	1			*194*	7	*301*	2	*147*	3	*215*	23	*180*	3	*120*	3		
	161	5					*303*	1					*184*	8	*122*	4		
	163	7											*202*	1	*124*	5		
	165	4																
RED SINDHI	*149*	5	*182*	28	*182*	23	*279*	4	*137*	20	*199*	5	*156*	2	*102*	12	*167*	4
	153	3	*184*	9	*184*	9	*287*	2	*139*	23	*209*	1	*160*	15	*104*	10	*169*	17
	155	4	*186*	34	*186*	9	*291*	1	*141*	2	*211*	12	*174*	13	*108*	17	*171*	4
	157	1	*192*	3	*190*	2	*293*	15	*143*	29	*215*	53	*176*	6	*110*	10	*175*	27
	161	9			*194*	33	*295*	28	*145*	1	*219*	1	*178*	4	*114*	1	*177*	21
	165	9					*297*	14	*149*	1			*180*	7	*120*	4	*179*	1
	167	41					*301*	6					*182*	1	*122*	1		
													184	25	*124*	19		
													188	1				

Appendix A.2 (cont.) *Frequencies of nine microsatellite loci in 7 Indian cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

Breed	Hel 9		Hel 13		ILSTS 005		ILSTS 006		INRA 005		INRA 023		INRA 032		INRA 035		INRA 063	
THARPARKER	*149*	13	*182*	39	*182*	3	*293*	21	*137*	8	*195*	5	*160*	9	*102*	16	*169*	2
	151	17	*186*	31	*184*	1	*295*	11	*139*	33	*197*	2	*174*	20	*104*	21	*175*	53
	155	1			*186*	19	*297*	6	*141*	7	*199*	1	*176*	26	*108*	4	*177*	11
	157	14			*190*	34	*299*	3	*143*	18	*205*	1	*180*	1	*110*	26		
	159	2			*194*	13	*301*	10			*211*	5	*184*	13	*120*	2		
	161	7					*305*	17			*215*	54	*188*	1	*122*	1		
	165	14																
	167	2																
DESI	*149*	10	*182*	23	*182*	9	*291*	1	*137*	18	*195*	6	*160*	9	*102*	16	*169*	14
	151	1	*184*	3	*184*	4	*293*	9	*139*	8	*197*	8	*174*	7	*104*	11	*173*	2
	155	9	*186*	25	*186*	16	*295*	19	*141*	14	*199*	10	*176*	11	*106*	1	*175*	30
	157	7	*192*	1	*190*	14	*297*	19	*143*	8	*203*	1	*178*	1	*108*	4	*177*	6
	161	6			*194*	9	*299*	1	*147*	4	*207*	1	*180*	3	*110*	7		
	163	6					*301*	1			*209*	1	*182*	2	*112*	2		
	165	8									*211*	2	*184*	16	*120*	4		
	167	3									*215*	23	*186*	1	*124*	6		
													188	2	*126*	1		
HARIANA	*149*	4	*182*	8	*184*	3	*283*	1	*137*	5	*195*	2	*160*	4	*104*	6	*169*	4
	151	3	*186*	12	*186*	8	*293*	5	*139*	3	*197*	4	*174*	3	*108*	2	*175*	13
	155	2			*190*	2	*295*	6	*141*	4	*199*	3	*176*	8	*110*	3	*177*	1
	157	3			*194*	7	*297*	7	*143*	7	*203*	1	*184*	2	*116*	2		
	161	2					*301*	1	*145*	1	*211*	2	*186*	1	*120*	5		
	163	1									*213*	2	*188*	2	*122*	1		
	165	2									*215*	5			*124*	1		
	167	3									*217*	1						

Appendix A.2 (cont.) *Frequencies of nine microsatellite loci in 7 Indian cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

Breed	Hel 9		Hel 13		ILSTS 005		ILSTS 006		INRA 005		INRA 023		INRA 032		INRA 035		INRA 063	
SAHIWAL	*149*	1	*182*	11	*184*	6	*293*	5	*137*	9	*197*	5	*160*	6	*100*	2	*167*	1
	151	3	*186*	9	*186*	9	*295*	3	*139*	7	*199*	1	*162*	1	*102*	1	*169*	7
	155	3			*190*	3	*297*	7	*141*	4	*205*	1	*164*	1	*104*	5	*175*	11
	157	6			*194*	2	*299*	1			*211*	1	*174*	9	*108*	1	*177*	1
	161	1					*301*	4			*215*	12	*184*	3	*110*	1		
	165	6													*116*	4		
															120	6		

Appendix A.3 *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
SOMBA	*98*	5	*123*	29	*118*	8	*144*	17	*141*	85	*162*	120	*142*	21	*180*	8	*128*	6	*183*	55
	104	2	*125*	3	*120*	11	*152*	103	*151*	30			*143*	1	*182*	104	*130*	102	*185*	60
	106	65	*131*	2	*122*	9			*153*	5			*144*	8	*184*	1	*132*	10	*189*	2
	108	8	*133*	3	*124*	18							*146*	6	*188*	3			*193*	1
	110	3	*135*	30	*126*	19							*148*	30						
	112	33	*139*	13	*128*	10							*150*	3						
			141	38	*130*	5							*152*	9						
			143	2	*132*	8							*154*	8						
					134	19							*156*	27						
					136	1							*158*	7						
					140	8														
LAGUNA	*100*	1	*123*	46	*120*	26	*144*	14	*141*	96	*162*	120	*142*	53	*180*	3	*128*	3	*181*	1
	104	1	*125*	19	*122*	13	*152*	106	*149*	3			*144*	2	*182*	116	*130*	113	*183*	69
	106	58	*131*	12	*124*	17			*151*	21			*148*	12	*184*	1	*132*	3	*185*	48
	108	3	*133*	4	*126*	16							*150*	4			*134*	1	*189*	1
	112	54	*135*	19	*128*	8							*152*	12					*193*	1
	160	1	*137*	7	*130*	15							*154*	6						
			139	5	*132*	4							*156*	31						
			141	4	*134*	12														
			143	4	*140*	3														

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
KAPSIKI	*98*	5	*123*	9	*118*	2	*144*	10	*141*	66	*162*	108	*142*	21	*180*	12	*128*	22	*183*	63
	100	1	*131*	17	*120*	27	*152*	110	*149*	20	*164*	4	*144*	1	*182*	108	*130*	90	*185*	39
	104	33	*134*	1	*122*	10			*151*	32			*148*	8			*132*	6	*189*	6
	106	51	*135*	28	*124*	9			*153*	2			*150*	7					*193*	12
	108	2	*137*	13	*126*	9							*152*	28						
	110	4	*139*	1	*128*	2							*154*	14						
	112	14	*141*	29	*130*	5							*156*	28						
	114	5	*143*	17	*132*	12							*158*	3						
	116	5	*145*	5	*134*	5							*160*	5						
					136	4							*164*	4						
					140	19							*165*	1						
KURI	*98*	4	*123*	9	*120*	48	*144*	17	*141*	45	*162*	87	*142*	14	*180*	8	*128*	24	*181*	6
	104	42	*131*	5	*122*	6	*152*	87	*149*	28	*164*	11	*144*	5	*182*	88	*130*	72	*182*	1
	106	31	*134*	1	*124*	13			*151*	28			*146*	6	*184*	1	*132*	6	*183*	31
	108	1	*135*	25	*126*	9			*153*	3			*148*	6	*186*	6			*185*	50
	110	5	*137*	15	*130*	4							*150*	7	*188*	1			*189*	8
	112	19	*139*	15	*132*	14							*152*	19					*193*	8
	118	2	*141*	18	*134*	5							*154*	12						
			143	10	*136*	1							*156*	10						
			145	6									*158*	20						
													160	1						
													162	1						
													164	3						

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
NAMCHI	*98*	12	*123*	27	*118*	8	*144*	17	*141*	80	*162*	100	*142*	11	*180*	16	*128*	12	*181*	2
	104	7	*131*	5	*120*	24	*152*	91	*149*	2			*144*	10	*182*	85	*130*	96	*183*	82
	106	43	*135*	9	*122*	23			*151*	24			*148*	4	*184*	3			*185*	17
	110	2	*137*	10	*124*	15			*153*	2			*150*	1	*186*	1			*189*	3
	112	38	*139*	19	*126*	2							*152*	19	*188*	1			*193*	4
	114	4	*141*	33	*128*	2							*154*	7						
			143	5	*130*	4							*156*	33						
					132	4							*158*	8						
					134	14							*160*	6						
					136	3							*162*	2						
					140	3							*166*	1						
MUTURU	*98*	1	*123*	11	*120*	23	*144*	17	*141*	22	*158*	1	*142*	25	*182*	55	*128*	2	*181*	1
	104	2	*131*	2	*122*	3	*152*	39	*149*	11	*162*	53	*148*	3	*188*	1	*130*	51	*183*	31
	106	36	*135*	21	*124*	21			*151*	18			*150*	1			*132*	3	*185*	19
	108	2	*137*	2	*126*	3			*153*	5			*152*	2					*189*	2
	110	2	*139*	6	*130*	2							*156*	20					*193*	1
	112	5	*141*	11	*132*	2							*158*	3						
	114	1	*143*	3									*162*	1						
	118	5											*164*	1						

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
BAOULE 1	*98*	10	*123*	45	*118*	6	*144*	25	*141*	52	*162*	88	*142*	8	*180*	1	*128*	1	*181*	8
	104	2	*127*	1	*120*	8	*152*	67	*149*	11			*144*	7	*182*	91	*130*	88	*183*	43
	106	24	*131*	6	*122*	5			*151*	26			*148*	14			*132*	1	*185*	36
	108	4	*135*	8	*124*	11			*153*	3			*150*	1			*134*	2	*189*	5
	110	1	*137*	6	*126*	22							*152*	15						
	112	47	*139*	14	*128*	7							*154*	8						
	114	2	*141*	6	*130*	2							*156*	17						
					132	8							*158*	21						
					134	15							*162*	1						
					136	6														
					138	2														
BAOULE 2	*98*	1	*123*	54	*118*	2	*144*	18	*141*	73	*158*	1	*142*	22	*180*	3	*128*	3	*183*	79
	104	5	*125*	4	*120*	23	*152*	92	*149*	6	*162*	109	*144*	3	*182*	105	*130*	106	*185*	29
	106	43	*135*	3	*122*	6			*151*	32			*148*	23			*132*	2	*187*	1
	108	7	*137*	1	*124*	21			*153*	1			*150*	4			*134*	1	*189*	1
	110	1	*139*	24	*126*	31							*152*	9						
	112	51	*141*	8	*128*	3							*154*	14						
	114	2	*143*	4	*130*	3							*156*	32						
					132	9							*158*	5						
					134	10														
					136	3														
					142	1														

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
BAOULE 3	*98*	4	*123*	58	*118*	1	*144*	14	*141*	70	*158*	2	*142*	20	*180*	15	*128*	3	*181*	1
	104	6	*131*	3	*120*	20	*152*	86	*149*	2	*162*	98	*144*	2	*182*	78	*130*	89	*183*	51
	106	47	*135*	5	*122*	6			*151*	27			*148*	25	*184*	5	*132*	4	*185*	34
	108	6	*137*	9	*124*	15			*153*	1			*150*	1			*134*	2	*193*	4
	110	1	*139*	17	*126*	20							*152*	9						
	112	29	*141*	1	*128*	9							*154*	4						
	114	3	*143*	1	*130*	5							*156*	35						
					132	9							*158*	3						
					134	4							*160*	1						
					136	10														
					140	1														
N'DAMA (MALI)	*98*	1	*123*	39	*118*	1	*144*	19	*141*	45	*158*	2	*142*	10	*180*	13	*128*	4	*181*	1
	104	7	*131*	3	*120*	17	*152*	69	*149*	15	*162*	86	*146*	1	*182*	64	*130*	75	*183*	55
	106	31	*135*	7	*122*	5			*151*	27			*148*	13	*184*	5	*132*	3	*185*	25
	108	1	*137*	2	*124*	10			*153*	1			*150*	4	*186*	3			*189*	1
	112	46	*139*	7	*126*	15							*152*	4	*188*	1			*193*	2
	116	1	*141*	19	*128*	10							*154*	9						
	118	1	*143*	4	*130*	5							*156*	25						
			145	1	*132*	4							*158*	6						
					134	10							*160*	10						
					136	2							*162*	5						
					138	9							*164*	1						

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
MBORORO	*98*	4	*123*	2	*120*	39	*144*	17	*141*	37	*162*	89	*142*	3	*180*	31	*128*	21	*181*	6
	104	39	*131*	19	*122*	5	*152*	77	*149*	21	*164*	1	*146*	1	*182*	39	*130*	59	*183*	23
	106	22	*133*	2	*124*	9			*151*	19			*148*	18	*184*	4	*132*	8	*185*	23
	110	5	*135*	19	*126*	16			*153*	8			*150*	9					*189*	26
	112	20	*137*	19	*128*	1			*161*	1			*152*	30					*193*	16
	116	3	*139*	6	*130*	3							*154*	9						
	118	1	*141*	11	*132*	7							*156*	11						
			143	10	*134*	4							*158*	9						
			147	6	*136*	5							*160*	2						
					140	3							*164*	1						
													166	1						
BORGOU	*98*	4	*123*	29	*118*	1	*144*	16	*141*	64	*162*	120	*142*	18	*180*	35	*128*	16	*181*	8
	104	34	*125*	4	*120*	37	*152*	104	*149*	7			*144*	4	*182*	75	*130*	96	*183*	32
	106	36	*131*	16	*122*	5			*151*	39			*146*	3	*184*	1	*132*	6	*185*	43
	108	8	*133*	1	*124*	19			*153*	4			*148*	20	*188*	1			*189*	25
	110	4	*135*	24	*126*	10							*150*	6					*193*	12
	112	31	*137*	16	*128*	14							*152*	21						
	114	1	*139*	9	*130*	10							*154*	12						
			141	10	*132*	10							*156*	26						
			143	8	*134*	8							*158*	4						
			147	3	*136*	3							*160*	1						
					140	1							*162*	3						
													164	1						
													168	1						

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
SOKOTO GUDALI	*98*	2	*123*	2	*118*	3	*144*	8	*141*	30	*162*	94	*142*	4	*180*	40	*128*	16	*181*	8
	104	49	*131*	16	*120*	33	*152*	86	*143*	2			*144*	1	*182*	48	*130*	59	*183*	19
	106	19	*135*	8	*122*	2			*149*	21			*146*	1	*184*	2	*132*	15	*185*	32
	108	1	*137*	23	*124*	6			*151*	29			*148*	30	*188*	4			*189*	27
	110	7	*139*	7	*126*	12			*153*	10			*150*	11					*193*	8
	112	13	*141*	14	*128*	4							*152*	15						
	114	2	*143*	17	*130*	6							*154*	7						
	118	1	*147*	7	*132*	5							*156*	19						
					134	15							*158*	1						
					136	7							*160*	1						
					140	1							*162*	2						
AZAWAK	*104*	21	*131*	4	*118*	4	*144*	5	*141*	9	*158*	1	*146*	4	*180*	10	*128*	4	*181*	21
	106	12	*135*	4	*120*	18	*152*	53	*149*	9	*162*	54	*150*	8	*182*	39	*130*	52	*183*	13
	110	20	*137*	9	*122*	2			*151*	17	*164*	1	*152*	12	*184*	2			*185*	9
	112	1	*139*	3	*124*	2			*153*	19			*154*	15	*188*	3			*189*	14
			141	7	*126*	2							*156*	4					*193*	1
			143	9	*128*	3							*158*	9						
					130	6							*160*	6						
					132	1														
					134	16														
					138	1														
					140	3														

Appendix A.3 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	HBB		BM2113		BoLA1		BoLA 2		RBP3		PRL		ETH131		His1		ILSTS014		ILSTS005	
PEUL FULANI	*98*	20	*123*	8	*118*	6	*144*	16	*141*	60	*162*	152	*142*	11	*180*	58	*128*	24	*181*	16
	104	47	*127*	3	*120*	54	*152*	142	*149*	21	*164*	4	*144*	11	*182*	82	*130*	120	*183*	59
	106	40	*131*	53	*122*	13			*151*	58			*146*	8	*184*	12	*132*	8	*185*	57
	108	1	*135*	3	*124*	10			*153*	19			*148*	19	*186*	4			*189*	18
	110	15	*137*	22	*126*	30							*150*	11	*188*	2			*193*	8
	112	20	*139*	2	*128*	5							*152*	40						
	118	11	*141*	24	*130*	10							*154*	22						
			143	30	*132*	6							*156*	16						
			145	3	*134*	7							*158*	13						
			146	2	*136*	3							*160*	3						
					138	2							*162*	4						
					140	12														
FULANI	*98*	5	*131*	6	*120*	16	*144*	11	*141*	21	*162*	48	*142*	1	*180*	21	*128*	12	*181*	5
	104	15	*135*	10	*122*	3	*152*	37	*149*	9			*148*	10	*182*	27	*130*	32	*183*	3
	106	14	*137*	12	*124*	3			*151*	13			*150*	2			*132*	4	*185*	15
	110	5	*139*	1	*126*	4			*153*	5			*152*	21					*189*	14
	112	8	*141*	7	*130*	6							*154*	3					*193*	11
	116	1	*143*	9	*132*	8							*155*	1						
			147	3	*134*	6							*156*	6						
					136	2							*158*	1						
													160	2						
													162	1						

Appendix A.4 *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BTMicros		HEL1		HEL5		OCAM		ETH152		ETH225		ILSTS001		RASA		TGLA48		TGLA116	
SOMBA	*143*	1	*103*	83	*165*	14	*184*	76	*193*	4	*140*	46	*77*	4	*182*	3	*73*	2	*79*	65
	157	1	*105*	16	*167*	11	*186*	42	*195*	1	*142*	1	*86*	6	*186*	41	*75*	57	*81*	49
	159	25	*107*	2	*169*	6	*190*	2	*197*	70	*144*	10	*87*	21	*188*	74	*77*	59	*83*	6
	161	76	*109*	19	*175*	2			*199*	43	*146*	13	*88*	16						
	163	11			*177*	74			*203*	2	*148*	25	*91*	1						
	167	6			*179*	13					*150*	16	*93*	65						
											152	2	*95*	5						
											158	7								
LAGUNA	*143*	1	*101*	5	*165*	26	*182*	1	*193*	5	*140*	11	*77*	3	*182*	2	*73*	12	*79*	76
	157	10	*103*	66	*167*	37	*184*	98	*197*	52	*144*	7	*87*	36	*186*	26	*75*	24	*81*	38
	159	4	*105*	6	*169*	2	*186*	21	*199*	58	*146*	1	*88*	2	*188*	92	*77*	84	*83*	6
	161	99	*107*	7	*175*	1			*201*	4	*148*	55	*91*	11						
	163	4	*109*	35	*177*	38			*203*	1	*150*	40	*93*	58						
	167	2	*117*	1	*179*	16					*158*	4	*95*	1						
													96	1						
KAPSIKI	*143*	15	*101*	2	*163*	9	*182*	15	*193*	31	*140*	20	*77*	23	*182*	25	*73*	27	*79*	59
	157	11	*103*	42	*165*	6	*184*	60	*197*	71	*144*	8	*87*	1	*186*	29	*75*	29	*81*	36
	159	25	*105*	26	*167*	22	*186*	44	*199*	11	*148*	47	*88*	34	*188*	60	*77*	62	*83*	25
	161	43	*107*	26	*171*	2	*190*	1	*203*	3	*150*	15	*91*	6	*190*	6				
	163	17	*109*	18	*175*	23					*158*	30	*93*	44						
	167	6	*111*	3	*177*	24							*95*	4						
	187	3	*113*	1	*179*	34							*96*	2						
			117	2																

Appendix A.4 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BTMicros		HEL1		HEL5		OCAM		ETH152		ETH225		ILSTS001		RASA		TGLA48		TGLA116	
KURI	*143*	18	*101*	7	*163*	9	*182*	34	*193*	61	*140*	10	*77*	36	*182*	11	*73*	30	*79*	46
	157	5	*103*	33	*165*	2	*184*	35	*197*	32	*144*	2	*87*	5	*186*	38	*75*	9	*81*	33
	159	26	*105*	18	*167*	25	*186*	35	*199*	4	*148*	17	*88*	24	*188*	34	*77*	65	*83*	24
	161	27	*107*	18	*171*	5			*201*	1	*150*	7	*91*	6	*190*	3			*91*	1
	163	16	*109*	16	*175*	5			*203*	4	*152*	3	*93*	25	*192*	3				
	165	3	*111*	1	*177*	32					*154*	1	*96*	2	*194*	15				
	167	9	*113*	3	*179*	26					*158*	59								
			117	8							*160*	1								
NAMCHI	*143*	2	*101*	2	*163*	6	*182*	8	*193*	30	*140*	26	*77*	19	*182*	9	*73*	9	*79*	56
	157	1	*103*	22	*165*	6	*184*	44	*195*	1	*146*	7	*86*	1	*186*	11	*75*	33	*81*	31
	159	31	*105*	15	*167*	28	*186*	54	*197*	61	*148*	55	*88*	30	*188*	63	*77*	64	*83*	20
	161	63	*107*	11	*169*	5			*199*	14	*150*	1	*91*	6	*190*	23			*85*	1
	163	6	*109*	56	*173*	1					*158*	15	*93*	52						
	165	1	*111*	2	*175*	3														
	167	2			*177*	19														
	187	2			*179*	38														
					181	2														
MUTURU	*157*	4	*103*	38	*163*	3	*182*	4	*193*	3	*140*	41	*77*	3	*182*	2	*73*	3	*79*	23
	159	3	*105*	13	*165*	5	*184*	26	*197*	33	*141*	1	*86*	2	*186*	18	*75*	39	*81*	32
	161	45	*107*	3	*167*	1	*186*	26	*199*	16	*144*	2	*88*	17	*188*	36	*77*	14	*83*	1
	163	1	*109*	2	*175*	4					*148*	10	*91*	5						
	167	1			*177*	17							*93*	25						
					179	22							*95*	1						
													97	3						

Appendix A.4 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BTMicros		HEL1		HEL5		OCAM		ETH152		ETH225		ILSTS001		RASA		TGLA48		TGLA116	
BAOULE 1	*153*	2	*103*	50	*161*	1	*182*	5	*193*	7	*140*	28	*77*	6	*182*	2	*73*	3	*79*	49
	159	17	*105*	20	*163*	4	*184*	67	*197*	44	*142*	2	*86*	6	*186*	15	*75*	57	*81*	29
	161	56	*107*	9	*165*	3	*186*	13	*199*	38	*144*	8	*87*	4	*188*	74	*77*	32	*83*	8
	163	7	*109*	13	*167*	12	*190*	3	*201*	1	*146*	2	*88*	1	*190*	1				
	165	6			*169*	4					*148*	15	*89*	1						
					173	1					*150*	13	*91*	13						
					177	39					*152*	13	*93*	28						
					179	12					*158*	9	*95*	3						
													96	6						
BAOULE 2	*153*	2	*103*	74	*165*	11	*184*	81	*193*	4	*140*	35	*77*	1	*182*	1	*73*	1	*79*	44
	159	37	*105*	10	*167*	28	*186*	16	*197*	44	*142*	4	*86*	5	*186*	36	*75*	64	*81*	54
	161	45	*107*	5	*169*	7	*190*	1	*199*	49	*144*	4	*87*	11	*188*	73	*77*	43		
	163	16	*109*	22	*175*	5			*201*	3	*146*	15	*88*	5						
	165	6	*111*	1	*177*	40					*148*	40	*91*	3						
	167	2			*179*	17					*150*	9	*93*	45						
											158	5	*95*	2						
													96	8						
BAOULE 3	*143*	2	*103*	57	*163*	1	*182*	3	*193*	20	*140*	43	*77*	11	*182*	2	*73*	3	*79*	32
	153	1	*105*	11	*165*	2	*184*	57	*197*	43	*142*	3	*86*	1	*186*	34	*75*	50	*81*	47
	159	32	*107*	12	*167*	33	*186*	30	*199*	35	*144*	2	*87*	6	*188*	60	*77*	45	*83*	7
	161	39	*109*	9	*169*	19	*190*	6			*148*	24	*88*	9	*190*	1				
	163	5	*111*	1	*175*	3					*150*	16	*91*	6	*192*	1				
	165	4	*117*	2	*177*	26					*152*	3	*93*	38						
	167	5			*179*	16					*158*	7	*95*	9						
	183	4																		

Appendix A.4 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BTMicros		HEL1		HEL5		OCAM		ETH152		ETH225		ILSTS001		RASA		TGLA48		TGLA116	
N'DAMA (MALI)	*143*	2	*101*	1	*163*	10	*182*	1	*193*	10	*140*	10	*77*	6	*182*	3	*73*	4	*79*	40
	159	24	*103*	41	*165*	13	*184*	41	*197*	46	*144*	13	*86*	1	*186*	31	*75*	49	*81*	29
	161	50	*105*	14	*167*	12	*186*	37	*199*	28	*148*	31	*87*	2	*188*	52	*77*	35	*83*	15
	163	5	*107*	10	*169*	4	*190*	5			*150*	24	*88*	10						
	167	1	*109*	17	*175*	6					*158*	8	*91*	1						
			111	3	*177*	26							*93*	36						
					179	13							*96*	14						
MBORORO	*143*	29	*99*	1	*163*	11	*182*	28	*193*	52	*140*	12	*77*	47	*182*	34	*73*	27	*79*	11
	145	2	*101*	7	*165*	5	*184*	52	*195*	2	*144*	1	*88*	8	*186*	36	*75*	33	*81*	38
	157	6	*103*	18	*167*	38	*186*	14	*197*	22	*148*	31	*91*	3	*188*	24	*77*	32	*83*	44
	159	12	*105*	15	*175*	1			*199*	5	*150*	1	*93*	32					*85*	1
	161	23	*107*	45	*177*	9			*201*	2	*154*	5	*96*	3						
	163	2	*109*	4	*179*	30			*205*	1	*158*	41	*97*	1						
	165	3	*111*	1							*160*	3								
	167	10	*117*	3																
	169	1																		
BORGOU	*143*	16	*101*	3	*161*	1	*182*	22	*193*	26	*140*	43	*77*	42	*182*	15	*73*	17	*79*	39
	145	3	*103*	53	*163*	9	*184*	68	*197*	54	*142*	3	*79*	1	*186*	44	*75*	43	*81*	43
	157	18	*105*	15	*165*	8	*186*	28	*199*	34	*144*	10	*87*	9	*188*	61	*77*	60	*83*	38
	159	22	*107*	34	*167*	22	*190*	2	*201*	2	*146*	4	*88*	11						
	161	36	*109*	12	*169*	11			*203*	4	*148*	16	*91*	4						
	163	17	*111*	2	*173*	2					*150*	7	*93*	48						
	167	6	*113*	1	*175*	6					*154*	1	*95*	1						
					177	35					*158*	36	*96*	2						
					179	18							*97*	2						

Appendix A.4 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BTMicros		HEL1		HEL5		OCAM		ETH152		ETH225		ILSTS001		RASA		TGLA48		TGLA116	
SOKOTO GUDALI	*143*	24	*101*	4	*163*	14	*182*	30	*193*	55	*140*	11	*77*	59	*182*	26	*73*	36	*79*	12
	157	6	*103*	32	*165*	1	*184*	48	*195*	1	*144*	4	*88*	1	*186*	37	*75*	19	*81*	47
	159	9	*105*	20	*167*	37	*186*	14	*197*	23	*148*	21	*91*	1	*188*	31	*77*	39	*83*	35
	161	32	*107*	28	*171*	1			*199*	10	*150*	6	*93*	29						
	163	12	*109*	5	*173*	2			*203*	2	*152*	2	*95*	4						
	165	3	*111*	2	*175*	3			*207*	1	*154*	2								
	167	6	*117*	3	*177*	10			*209*	2	*158*	48								
	169	1			*179*	20														
	187	1																		
AZAWAK	*143*	6	*99*	2	*163*	6	*182*	12	*193*	40	*140*	1	*77*	44	*182*	5	*73*	25	*79*	6
	157	2	*101*	4	*165*	3	*184*	31	*197*	7	*144*	2	*87*	2	*186*	33	*75*	19	*81*	16
	159	18	*103*	12	*167*	21	*186*	13	*199*	2	*148*	19	*93*	12	*188*	20	*77*	12	*83*	32
	161	5	*105*	4	*177*	6			*201*	7	*150*	2								
	163	5	*107*	17	*179*	20					*152*	12								
	167	17	*109*	11							*158*	22								
	187	5																		
PEUL FULANI	*143*	43	*101*	4	*161*	1	*182*	41	*193*	97	*140*	36	*77*	70	*182*	41	*71*	1	*79*	14
	157	6	*103*	44	*163*	15	*184*	67	*197*	28	*144*	10	*88*	3	*186*	69	*73*	47	*81*	61
	159	24	*105*	35	*165*	10	*186*	16	*199*	4	*148*	22	*91*	7	*188*	45	*75*	52	*83*	65
	161	34	*107*	44	*167*	27			*203*	3	*150*	6	*93*	8	*190*	1	*77*	50	*85*	2
	163	15	*109*	11	*169*	8					*152*	2	*96*	6						
	165	11	*117*	12	*171*	2					*154*	3								
	167	18			*177*	24					*156*	1								
	187	7			*179*	57					*158*	77								
											160	1								

Appendix A.4 (cont.) *Frequencies of ten microsatellite loci in West African cattle breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BTMicros		HEL1		HEL5		OCAM		ETH152		ETH225		ILSTS001		RASA		TGLA48		TGLA116	
FULANI	*143*	15	*101*	6	*163*	10	*182*	11	*193*	30	*140*	14	*77*	31	*182*	12	*73*	13	*79*	11
	157	3	*103*	11	*167*	14	*184*	25	*195*	2	*144*	1	*88*	1	*186*	17	*75*	14	*81*	22
	159	6	*105*	7	*175*	2	*186*	11	*197*	13	*148*	7	*91*	2	*188*	19	*77*	21	*83*	15
	161	14	*107*	19	*177*	7	*190*	1	*199*	1	*150*	3	*93*	12						
	165	1	*109*	4	*179*	15			*201*	1	*158*	23	*96*	2						
	167	7	*111*	1					*203*	1										
	187	2																		

Appendix A.5 *Frequencies of ten X-chromosome microsatellite loci typed in 22 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BL1045		BMS1616		XBM361		UWCA19		BMS2713		BL1098		BMS960		XBM7		XBM701		BMS649	
ABERDEEN ANGUS	*94*	29	*94*	11	*158*	28	*101*	3	*153*	29	*219*	2	*110*	1	*192*	14	*148*	3	*120*	4
			100	3			*107*	26			*223*	2	*112*	11	*194*	2	*150*	1	*122*	22
			104	10							*225*	2	*114*	16	*196*	9	*152*	25	*124*	2
			106	3							*227*	23			*198*	4				
CHAROLAIS	*94*	32	*94*	10	*156*	3	*107*	33	*153*	33	*213*	1	*110*	2	*192*	13	*152*	33	*120*	6
			96	1	*158*	28					*219*	1	*112*	17	*194*	4			*122*	25
			104	18	*160*	1					*223*	2	*114*	11	*196*	15			*124*	1
			108	4							*225*	3	*120*	1	*198*	1			*126*	1
											227	25	*128*	1						
											231	1								
N'DAMA (GUINEA)	*94*	24	*94*	20	*154*	9	*107*	35	*151*	22	*213*	6	*110*	1	*186*	11	*152*	35	*116*	1
	100	12	*96*	1	*156*	2			*153*	14	*219*	5	*112*	15	*188*	8			*120*	16
			100	6	*158*	16					*221*	1	*114*	5	*190*	1			*122*	16
			104	1	*160*	3					*227*	23	*118*	3	*192*	11			*126*	1
			106	6									*120*	2	*194*	2			*128*	1
			108	2									*122*	10	*196*	3				
N'DAMA (GUINEA BISSAU)	*94*	8	*94*	11	*154*	3	*107*	21	*151*	5	*203*	2	*110*	1	*186*	4	*152*	22	*120*	8
	100	14	*100*	6	*158*	14			*153*	17	*209*	6	*112*	5	*192*	17			*122*	13
			104	2	*160*	3					*215*	4	*114*	2	*194*	1			*128*	1
			106	3	*162*	1					*217*	9	*118*	2						
											219	1	*120*	2						
													122	10						

Appendix A.5 (cont.) *Frequencies of ten X-chromosome microsatellite loci typed in 22 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BL1045		BMS1616		XBM361		UWCA19		BMS2713		BL1098		BMS960		XBM7		XBM701		BMS649	
SOMBA	*94*	8	*94*	5	*154*	3	*107*	10	*143*	1	*219*	1	*112*	1	*188*	5	*152*	9	*120*	6
	100	2	*100*	3	*156*	1			*151*	6	*223*	1	*114*	4	*192*	1	*162*	1	*122*	4
			104	1	*158*	3			*153*	3	*227*	8	*118*	1	*194*	4				
					160	2							*122*	4						
GOBRA	*92*	5	*92*	1	*150*	4	*105*	21	*143*	15	*203*	1	*106*	3	*186*	5	*152*	12	*120*	19
	94	9	*94*	8	*152*	10	*107*	11	*151*	5	*207*	1	*108*	2	*188*	3	*158*	5	*122*	8
	98	16	*96*	11	*154*	3			*153*	12	*209*	3	*110*	8	*190*	1	*160*	9	*126*	1
	100	2	*100*	6	*156*	1					*215*	3	*112*	1	*192*	12	*162*	4	*130*	2
			104	2	*158*	9					*217*	6	*114*	8	*194*	6	*164*	1	*132*	2
			106	3	*160*	4					*219*	8	*118*	4	*200*	5				
			112	1							*223*	4	*120*	1						
											225	5	*122*	5						
											227	1								
SOKOTO GUDALI	*92*	1	*92*	1	*150*	2	*105*	10	*143*	8	*209*	2	*106*	1	*188*	2	*152*	1	*120*	4
	94	1	*94*	1	*152*	6	*107*	3	*153*	5	*215*	1	*108*	3	*192*	5	*158*	3	*130*	3
	98	9	*96*	8	*158*	5					*217*	1	*114*	7	*194*	2	*160*	2	*132*	6
	100	2	*112*	1							*219*	5	*118*	2	*196*	2	*164*	7		
			114	1							*225*	3			*200*	2				
			116	1							*227*	1								
MBORORO	*92*	2	*94*	2	*152*	4	*105*	6	*143*	6	*209*	1	*108*	2	*188*	2	*152*	2	*120*	3
	94	2	*96*	4	*154*	1	*107*	2	*153*	3	*215*	1	*110*	2	*190*	1	*158*	2	*122*	2
	98	5	*104*	1	*158*	2					*219*	6	*112*	1	*192*	4	*160*	2	*130*	1
			114	1							*227*	1	*114*	3			*164*	3	*132*	3
													122	1						

Appendix A.5 (cont.) *Frequencies of ten X-chromosome microsatellite loci typed in 22 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BL1045		BMS1616		XBM361		UWCA19		BMS2713		BL1098		BMS960		XBM7		XBM701		BMS649	
KURI	*94*	8	*94*	2	*152*	3	*105*	9	*143*	4	*209*	2	*106*	2	*180*	1	*152*	7	*120*	5
	98	4	*96*	2	*154*	1	*107*	4	*153*	7	*215*	2	*108*	1	*186*	1	*158*	1	*122*	5
			100	3	*156*	2			*155*	2	*217*	5	*110*	3	*188*	1	*160*	3	*128*	1
			104	3	*158*	7					*219*	2	*112*	1	*192*	7	*164*	2	*130*	2
			106	3							*225*	1	*114*	1	*194*	1				
											227	1	*120*	1	*196*	1				
													122	4						
ANATOLIAN BLACK	*94*	11	*94*	1	*158*	12	*105*	1	*153*	11	*217*	1	*106*	1	*192*	2	*152*	8	*118*	1
	98	1	*96*	4			*107*	11			*225*	1	*112*	8	*196*	10	*160*	2	*120*	3
			102	1							*227*	7	*114*	3					*122*	5
			104	4							*229*	1							*126*	1
			112	1															*130*	1
																			134	1
IRAQI	*92*	1	*96*	6	*150*	4	*105*	10	*143*	14	*215*	1	*110*	3	*188*	1	*152*	4	*116*	1
	94	7	*104*	1	*152*	10	*107*	4	*153*	1	*217*	1	*112*	4	*192*	7	*156*	1	*120*	2
	98	6	*106*	2	*158*	2					*221*	4	*114*	6	*196*	1	*160*	2	*122*	3
	100	2	*114*	6							*223*	2	*118*	1	*198*	1	*162*	1	*130*	1
											227	1			*200*	1	*164*	3	*132*	6
											229	1			*202*	4	*166*	2	*134*	2
											235	6					*168*	2		

Appendix A.5 (cont.) *Frequencies of ten X-chromosome microsatellite loci typed in 22 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BL1045		BMS1616		XBM361		UWCA19		BMS2713		BL1098		BMS960		XBM7		XBM701		BMS649	
NELLORE	*92*	3	*96*	1	*150*	1	*105*	11	*143*	11	*225*	6	*106*	2	*188*	2	*152*	3	*122*	2
	94	1	*104*	1	*152*	7	*107*	1	*153*	1	*227*	4	*108*	3	*192*	3	*160*	4	*128*	1
	98	8	*110*	2	*158*	2					*229*	1	*110*	3	*194*	1	*164*	5	*130*	7
			112	4									*112*	1	*196*	2			*132*	2
			114	2									*114*	3	*198*	3				
															202	1				
ONGOLE	*92*	2	*92*	1	*150*	1	*105*	3	*143*	3	*225*	1	*106*	1	*188*	1	*158*	1	*118*	1
	98	1	*96*	2	*152*	2					*229*	2	*114*	2	*192*	1	*160*	2	*128*	1
															196	1			*132*	1
AFRIKANER	*98*	13	*92*	6	*154*	1	*105*	5	*151*	2	*205*	2	*110*	6	*188*	2	*152*	3	*122*	11
	102	1	*94*	1	*158*	5	*107*	3	*153*	8	*207*	4	*112*	1	*190*	1	*154*	1	*124*	1
			96	2							*209*	1	*114*	2	*192*	2	*162*	7	*132*	2
			100	2							*211*	2	*124*	1	*194*	3				
			104	2							*215*	1			*196*	1				
			108	1							*217*	1								
											219	1								
ANKOLE	*96*	3	*92*	1	*142*	9	*105*	7	*143*	9	*207*	1	*108*	1	*190*	1	*162*	6	*128*	1
	98	1	*94*	3	*154*	1			*153*	1	*211*	1	*110*	4	*192*	1	*164*	2	*130*	1
	102	4	*104*	1							*213*	1	*112*	2	*194*	3	*166*	3	*132*	11
			112	4							*215*	3	*114*	5	*202*	3				
											219	1								
											221	1								
											225	1								

Appendix A.5 (cont.) *Frequencies of ten X-chromosome microsatellite loci typed in 22 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BL1045		BMS1616		XBM361		UWCA19		BMS2713		BL1098		BMS960		XBM7		XBM701		BMS649	
ARASHIE	*96*	5	*94*	1	*142*	14	*105*	10	*143*	14	*215*	2	*106*	2	*188*	1	*152*	2	*120*	1
	98	4	*96*	7	*152*	2	*107*	1	*151*	2	*217*	1	*110*	4	*192*	3	*160*	2	*122*	5
	102	3	*104*	1							*219*	2	*112*	1	*194*	2	*162*	3	*124*	1
	104	3	*112*	4							*221*	4	*114*	3	*196*	4	*166*	7	*132*	6
	106	1									*223*	1	*122*	1	*202*	2			*134*	2
											225	2							*136*	1
											227	2								
BAROTSE	*98*	4	*92*	1	*152*	3	*107*	16	*151*	2	*205*	3	*110*	3	*186*	3	*152*	16	*120*	1
	100	5	*94*	1	*154*	6			*153*	6	*207*	2	*112*	6	*188*	2			*122*	8
	104	4	*96*	2					*155*	1	*209*	8	*114*	3	*192*	9			*124*	3
			98	1							*213*	1	*122*	2	*194*	1			*130*	1
			100	2							*217*	1							*132*	1
			104	1							*227*	1								
			108	1																
			110	1																
			112	1																
			118	2																
KAVANGO	*98*	6	*92*	1	*152*	5	*107*	11	*151*	5	*207*	1	*110*	4	*186*	7	*152*	12	*118*	1
	104	2	*94*	3	*154*	6			*153*	6	*209*	11	*112*	6	*192*	2			*122*	4
			100	2									*124*	1					*132*	2
			104	2															*134*	1
			106	1															*140*	1
			108	1																
			112	1																

Appendix A.5 (cont.) *Frequencies of ten X-chromosome microsatellite loci typed in 22 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BL1045		BMS1616		XBM361		UWCA19		BMS2713		BL1098		BMS960		XBM7		XBM701		BMS649	
NGUNI	*98*	5	*94*	4	*152*	1	*107*	9	*151*	1	*209*	5	*110*	3	*186*	1	*148*	1	*122*	7
	100	1	*100*	1	*154*	5			*153*	5	*211*	4	*112*	3	*188*	3	*150*	1	*130*	1
	102	1	*104*	1									*114*	1	*192*	1	*152*	6	*132*	1
			106	2									*124*	2	*194*	2	*154*	1		
															196	2				
OGADEN	*96*	2	*92*	1	*142*	16	*105*	10	*143*	16	*213*	4	*106*	2	*190*	1	*152*	4	*122*	2
	98	1	*94*	6	*152*	1	*107*	7	*151*	1	*219*	4	*108*	1	*192*	4	*160*	4	*132*	9
	102	14	*106*	3	*154*	1			*153*	1	*223*	5	*110*	3	*194*	7	*162*	7	*142*	2
	104	1	*110*	1							*225*	1	*114*	8	*198*	2	*166*	3		
			112	3							*227*	1			*200*	2				
			118	1																
WATUSI	*98*	2	*92*	3	*154*	3	*107*	2	*153*	3	*211*	1	*112*	2	*192*	1	*152*	2	*122*	3
	102	1									*215*	1	*114*	1	*194*	1	*162*	1		
															196	1				
ZEBU MALAGASY	*100*	1	*104*	2	*142*	2	*105*	1	*143*	2	*209*	1	*112*	3	*188*	1	*152*	3	*132*	4
	104	1	*114*	2	*152*	1	*107*	3	*151*	1	*213*	1	*114*	1	*194*	1	*162*	1		
	106	1									*215*	1			*202*	2				
											221	1								

Appendix A.6 *Frequencies of 5 X-chromosome microsatellite loci typed in 13 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BMS 811		XBM 521		XBM111		BMS1820		ED1MS1	
ABERDEEN ANGUS	*111*	19	*125*	9	*133*	5	*101*	12	*131*	27
	113	9	*127*	20	*135*	24	*103*	17	*135*	1
	115	1								
CHAROLAIS	*107*	1	*125*	5	*133*	12	*99*	1	*115*	2
	111	17	*127*	27	*135*	20	*101*	12	*123*	1
	113	15	*131*	1	*137*	1	*103*	19	*131*	25
									133	4
N'DAMA (GUINEA)	*111*	22	*125*	32	*129*	1	*101*	14	*109*	1
	113	3	*127*	3	*133*	4	*103*	21	*115*	20
	115	11			*135*	31			*123*	1
									127	9
									129	1
									131	2
									133	2
N'DAMA (GUINEA BISSAU)	*107*	1	*125*	16	*129*	6	*101*	12	*115*	14
	109	1	*127*	6	*135*	16	*103*	10	*127*	1
	111	12							*131*	1
	113	1							*133*	6
	115	7								

Appendix A.6 (cont.) *Frequencies of 5 X-chromosome microsatellite loci typed in 13 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BMS 811		XBM 521		XBM111		BMS1820		ED1MS1	
SOMBA	*111*	9	*125*	5	*129*	3	*101*	3	*115*	3
	115	1	*127*	4	*135*	7	*103*	7	*117*	1
			129	1					*127*	1
									129	2
									131	1
									133	1
									139	1
GOBRA	*103*	6	*125*	11	*129*	15	*101*	28	*115*	6
	107	1	*127*	7	*133*	2	*103*	4	*117*	3
	109	1	*129*	13	*135*	15			*119*	1
	111	17							*121*	5
	113	2							*125*	10
	115	4							*129*	3
									133	1
									135	2
SOKOTO GUDALI	*103*	5	*125*	5	*129*	7	*101*	11	*115*	1
	105	1	*129*	8	*135*	6	*103*	2	*117*	3
	109	1							*121*	4
	111	6							*125*	2
									127	2
									131	1

Appendix A.6 (cont.) *Frequencies of 5 X-chromosome microsatellite loci typed in 13 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BMS 811		XBM 521		XBM111		BMS1820		ED1MS1	
MBORORO	*103*	2	*125*	4	*129*	3	*101*	8	*115*	2
	111	5	*127*	1	*135*	6	*103*	1	*117*	2
	113	1	*129*	3					*121*	1
									125	3
									129	1
KURI	*103*	3	*125*	7	*129*	4	*99*	1	*111*	1
	109	1	*127*	3	*135*	9	*101*	7	*115*	3
	111	7	*129*	3			*103*	4	*117*	2
	115	2							*121*	1
									125	2
									127	1
									129	1
ANATOLIAN BLACK	*103*	1	*125*	6	*133*	6	*101*	1	*115*	3
	111	2	*127*	6	*135*	6	*103*	11	*119*	1
	113	5							*131*	5
	115	1							*135*	2
	117	1								
	119	1								
	121	1								

Appendix A.6 (cont.) *Frequencies of 5 X-chromosome microsatellite loci typed in 13 breeds. Allele sizes are shown in italics and the number of individuals in regular type.*

BREED	BMS 811		XBM 521		XBM111		BMS1820		ED1MS1	
IRAQI	*103*	2	*125*	8	*129*	13	*101*	3	*115*	1
	111	4	*127*	4	*135*	3	*103*	12	*121*	1
	113	8	*129*	4					*125*	7
									131	4
									133	1
									135	2
NELLORE	*103*	4	*125*	9	*129*	10	*101*	7	*113*	1
	113	3	*129*	3	*135*	2	*103*	5	*125*	9
	115	1							*129*	2
	117	4								
ONGOLE	*103*	1	*125*	3	*129*	3	*101*	3	*121*	1
	113	1							*125*	1
	117	1							*129*	1

APPENDIX B

BREED CODES

Appendix B *Table of domestic cattle breeds and abbreviated codes used in this thesis.*

BREED	CODE	BREED	CODE	BREED	CODE
Aberdeen Angus	**AN**	*Hereford*	**HE**	*N'Dama Nigeria*	**NDN**
Abigar	**AI**	*Highland*	**HI**	*N'Dama Senegal*	**NDS**
Afar	**AFA**	*HighlandZebu*	**HZ**	*Nellore*	**NE**
Afrikaner	**AF**	*Holstein*	**HL**	*Nguni*	**NG**
Alentejana	**AL**	*Horro*	**HR**	*Nkone*	**NK**
Anatolian Black	**AB**	*Hungarian Gray*	**HG**	*Nuba*	**NU**
Angoni	**AG**	*Iraqi*	**IR**	*Ogaden*	**OG**
Ankole	**AK**	*Iringared*	**IRG**	*Ongole*	**ON**
Arado	**AO**	*Jersey*	**JE**	*Ormaboran*	**OR**
ArashieBeja	**AS**	*Kapsiki*	**KP**	*Pedi*	**PE**
Arouquesa	**AQ**	*Kavango*	**KV**	*Peul Fulani*	**PF**
Arsi	**ARS**	*Kavirondo*	**KVR**	*Pinsgaur*	**PS**
Azawak	**AZ**	*Kenana*	**KE**	*Pustertaler*	**PT**
Baladi	**BAL**	*Kenyan Boran*	**KB**	*Raya-azebo*	**RZ**
Bale	**BA**	*Kerry*	**KR**	*Red Sindhi*	**RS**
Baoule	**BL**	*Kigezi*	**KG**	*Retinta*	**RE**
Barotse	**BA**	*Kilimanjaro*	**KIL**	*Romagnola*	**RO**
Blonde	**BD**	*Koakoland*	**KO**	*Sahiwal*	**SA**
Boran Ethopia	**BE**	*Kurdi*	**KUR**	*Saler*	**SL**
Borgou	**BO**	*Kuri*	**KU**	*Sheko*	**SH**
Butana	**BU**	*Laguna*	**LA**	*Simmenthal*	**SI**
ButanaFoya	**BF**	*Landim*	**LD**	*Sokoto Gudali*	**SG**
Charolais	**CH**	*Malawizebu*	**MZ**	*Somba*	**SO**
Damascus	**DM**	*Marmarena*	**MM**	*South Anatolian Red*	**SAR**
Danakil	**DN**	*Mashona*	**MS**	*Swiss Brown*	**OB**
Desi	**DE**	*Maure*	**MA**	*Sykia*	**SK**
East Anatolian Red	**EAR**	*Mbororo*	**MB**	*Tharparker*	**TH**
Egyptian	**EG**	*Mertolenga*	**MT**	*Tonga*	**TO**
Eringer	**ER**	*Modicana*	**MD**	*Tuli*	**TU**
Evolenard	**EV**	*Muturu*	**MU**	*Turkish Gray*	**TG**
Fogera	**FO**	*Namchi*	**NA**	*Vosges*	**VO**
Friesian	**FR**	*N'Dama Bissau*	**NB**	*Watusi*	**WA**
Fulani	**FU**	*N'Dama Gambia*	**NGa**	*White Fulani*	**WF**
Gobra	**GO**	*N'Dama Guinea*	**NGu**	*ZebuMalagasy*	**ZM**
Hariana	**HA**	*N'Dama Mali*	**NDM**		

APPENDIX C

SOLUTIONS

Appendix C *Buffers and Solutions*

Platinum *Taq* DNA polymerase buffer (GIBCO BRL)

200 mM	Tris-HCl (*pH* 8.4),
500 mM	KCl

10 X TBE electrophoresis buffer (INVITROGEN)

100 M	Tris-borate (*pH* 8.3)
0.02 M	EDTA (*pH* 8.0)

100 X TE Buffer (SIGMA-ALDRICH)

1.0 M	Tris-Cl (*pH* 8.0)
0.1 M	EDTA

6% denaturing polyacrylamide gel (50mls)

5.0 mls	Sequagel™buffer
12.0 mls	Sequagel™concentrate
33.0 mls	Sequagel™diluent

Gel components were supplied by NATIONAL DIAGNOSTICS.

Final concentrations were: acrylamide:bisacrylamide (19:1), 5% Urea and 1XTBE.
Gel was polymerised with:

200 µls	10% Ammonium Persulphate (APS)
20 µls	N, N, N' N'- Tetramethylethylenediamine TEMED

5% Non-denaturing polyacrylamide gel (50mls)

6.25 mls	Accugel™ solution
2.5 mls	Glycerol (100%)
5.0 mls	10X TBE buffer
36.25mls	ddH_20

Gel components were supplied by NATIONAL DIAGNOSTICS.

Final concentrations were acrylamide:bisacrylamide (29:1), 1x TBE buffer and 5% glycerol.
Gel was polymerised with:
200 µls 10% APS
20 µl TEMED