Edith Cowan University

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study.

The University does not authorize you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following:

- Copyright owners are entitled to take legal action against persons who infringe their copyright.

- A reproduction of material that is protected by copyright may be a copyright infringement. Where the reproduction of such material is done without attribution of authorship, with false attribution of authorship or the authorship is treated in a derogatory manner, this may be a breach of the author's moral rights contained in Part IX of the Copyright Act 1968 (Cth).

- Courts have the power to impose a wide range of civil and criminal sanctions for infringement of copyright, infringement of moral rights and other offences under the Copyright Act 1968 (Cth). Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

THE GENETIC HERITAGE OF CHINA

A genomic study of PR China based on nine representative ethnic populations

Michael Black

BBiotech, MSc (Human Biology)

This thesis is presented in fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Faculty of Computing, Health and Science
Edith Cowan University

July 2005

USE OF THESIS

The Use of Thesis statement is not included in this version of the thesis.

ABSTRACT

During the course of the last decade, genetic data have increasingly complemented linguistic, archaeological and palaeontological evidence in efforts to reconstruct human history. As technology has developed, studies have utilised genomic techniques in tracing the origins and migratory patterns of modern humans. East Asia is a particular hotspot of human migration, especially Mainland China where a large number of human fossils have been unearthed and more than 20% of the world's population now resides.

There are 56 officially recognised ethnic groups (*minzu*) within the population of PR China which totals 1,300 million. The majority Han population is distributed throughout the country and forms 90% of the total, whereas the other 55 minority populations mostly live in peripheral and boundary regions. To date, information on these minorities has been fragmentary and, from both evolutionary and historical perspectives, data on their genetic profiles would be of considerable value in identifying their founding populations and genetic inter-relationships. There are also strongly conflicting opinions on the origins of the Han and the degree to which they can be regarded as genetically homogenous.

The current study measured the genetic diversity and ancestry of nine ethnic populations resident in PR China. In addition to the Han, these study populations comprised the Miao, Yao, Kucong and Tibetan communities from Yunnan province in the southwest of the country, and four Muslim populations, the Hui, Bo'an, Dongxiang and Sala from northern and central China. Both biparental and uniparental genetic influences on the populations were examined by the analysis of autosomal, mitochondrial and Y-chromosome markers.

In general, it was found that the study populations displayed diverse paternal ancestries but more restricted maternal ancestries. From the Y-chromosome data in particular, major events such as the Neolithic population expansion and more recent historical events, such as migration along the Silk Road, could be inferred. Through the use of autosomal markers, aspects of the internal structure of the study populations were uncovered, such as endogamy and/or consanguinity. These conclusions were made possible, in part, by experimental Likelihood-based stochastic coalescent modelling.

Intriguingly, it was revealed that the Kucong of Yunnan, an ethnic group not previously surveyed for genetic diversity and not accorded official minority status within PR China, could possibly be representative of indigenous populations dating from the first migrations into East Asia.

While other more recent events could be inferred from summary statistics and phylogenetic and coalescent-based genetic analyses of the study populations, the changing definition of the ethnic study populations themselves proved to be the most important factor. It is therefore recommended that future studies primarily utilize a community-by-community approach, and not rely on the official *minzu* category as an accurate indicator of genetic ancestry.

ACKNOWLEDGEMENTS

Firstly, I would like to especially thank my principal supervisor Professor Alan Bittles for his boundless enthusiasm and faith in my abilities and, most importantly, sharing the contents of his cellar. I would also like to thank my family for supporting and encouraging me during the long years of study.

I would like to thank the 'Pax 7 Gang' for their support, encouragement and especially the lunches. Finally I would like to thank those involved in the collection of samples used in this study, especially my former supervisor Dr Wei Wang, without whom this study would not be possible.

TABLE OF CONTENTS

ABSTRACT iii

DECLARATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF FIGURES i

LIST OF EQUATIONS xvi

LIST OF TABLES xviii

CHAPTER 1 INTRODUCTION 1

1.1 Introduction 2

1.2 Significance of the study 5

1.3 Aims of the Study 6

1.4 Research Issues and Questions 7

CHAPTER 2 REVIEW OF MOLECULAR ANTHROPOLOGY 8

2.1 Molecular anthropology and the search for human origins 9

2.2 Molecular systems in molecular anthropology 11

2.2.1 The human mitochondrial genome 11

2.2.2 Genetic diversity of the human Y-chromosome 13

2.2.3 Autosomal and X-Chromosome Approaches 16

2.3 Molecular anthropology and PR China 17

CHAPTER 3 REVIEW OF POPULATION GENETIC THEORY 21

3.1 Introduction 22

3.2 The Neutral Theory 22

3.3 The Wright-Fisher Neutral Model 23

3.3.1 Summary Statistics under the WFN model 25

3.3.2 Population differentiation 26

3.3.2.1 Wright's F_{ST} model 27

3.3.2.2 Weir and Cockerham's approach 27

3.3.2.3 Nei's G_{ST} 28

3.3.2.4 Analysis of molecular variance 29

3.4 Phylogenetic Methods 32

3.4.1 Distance matrix methods 32

3.4.1.1 Drawing the trees: the Neighbour-Joining algorithm 34

3.4.1.2 Ordination methods 36

3.4.2 Discrete Methods 38

3.4.2.1 Maximum Parsimony 38

3.4.2.2 Maximum Likelihood 38

3.4.3 Network Analysis 39

3.5 Coalescent Theory 41

3.5.1 Likelihood analysis and the Coalescent Theory 43

3.5.2 Markov Chain Monte Carlo 44

3.5.3 Maximum Likelihood estimation of parameters using MCMC 47

3.5.4 Bayesian analysis and the Coalescent Theory 48

3.5.5 Bayesian population assignment 50

3.6 Summary 54

CHAPTER 4 SUBJECTS AND METHODS 55

4.1 Introduction 56

4.2 Historical background of the study populations 56

4.2.1 The 'Ancient' populations 59

4.2.1.1 Hanzu 59

4.2.1.2 Kucong 59

4.2.1.3 Miaozu 59

4.2.1.4 Yaozu 60

4.1.2.5 Zangzu (Tibetan) 61

4.2.2 The 'Recent' populations 62

4.2.2.1 Bo'anzu 62

4.2.2.2 Dongxiangzu 63

4.2.2.3 Huizu 63

4.2.2.4 Salazu 64

4.3 Sample collection 65

4.4 Extraction of DNA from blood spots 66

4.5 Quantitation of DNA concentration 66

4.6 Polymerase Chain Reaction 67

4.7 Autosomal STR analysis 67

4.7.1 Autosomal STR PCR conditions 69

4.8 Agarose gel electrophoresis of microsatellite sequences 70

4.9 Resolution of STR sequence size through fluorescent labelling 71

4.10 Detection and quantification of STR markers 71

4.11 Y-chromosome analysis 72

4.11.1 Y-chromosome STR PCR conditions 75

4.11.2 Y-chromosome biallelic polymorphism PCR 76

4.11.3 YAP and RFLP assays 76

4.12 Mitochondrial DNA analysis 78

4.12.1 HV1 sequencing 78

4.12.2 mtDNA 9-bp deletion detection 81

4.12.3 mtDNA haplogrouping protocol 81

CHAPTER 5 SUMMARY STATISTIC ANALYSIS 84

5.1 Introduction 85

5.2 Autosomal microsatellite analysis 85

5.2.1 Autosomal summary statistics 85

5.2.2 Autosomal AMOVA analysis 89

5.2.3 Autosomal phylogenetic analysis 92

5.3 Y-chromosome microsatellite analysis 95

5.3.1 Y-chromosome microsatellite summary statistics 95

5.3.2 Haplotype analysis 100

5.3.3 Y-chromosome microsatellite AMOVA analysis 101

5.3.4 Y-chromosome phylogenetic analysis 102

5.4 Mitochondrial HV1 sequence analysis 106

5.4.1 mtDNA summary statistics 106

5.4.2 mtDNA mismatch distributions 109

5.4.3 Maternal genetic structure 110

5.4.4 mtDNA phylogenetic analysis 110

5.5 Summary 114

CHAPTER 6 GENEALOGICAL HAPLOGROUP ANALYSIS 115

6.1 Introduction 116

6.2 Mitochondrial haplogrouping 116

6.2.1 Principal Components Analysis of mtDNA haplogroups 119

6.3 Y-chromosome haplogroups 122

6.3.1 Haplotype diversity in the 'Ancient' populations 122

6.3.2 Haplogroup diversity in the 'Recent' populations 125

6.3.3 Principal Components Analysis of Y-chromosome haplogroups 127

6.3.4 Network analysis 127

6.3.4.1 Maximum Parsimony Median Network Diagrams 129

6.3.4.2 Rho statistics and TMRCA estimates 131

6.3.5 Combining inter- and intra-population structures 135

6.4 Summary 139

CHAPTER 7 LIKELIHOOD-BASED STOCHASTIC COALESCENT MODELLING 140

7.1 Introduction 141

7.2 Maximum-Likelihood Estimation of Genetic Diversity 142

7.3 Autosomal genotypic structure 151

7.3.1 Inter-population genotypic assignment 152

7.3.2 Intra-population genotypic assignment 154

7.4 Bayesian inference of paternal ancestry 158

7.4.1 Inter-population analysis 159

7.4.2 Batwing intra-population analysis 166

7.4.2.1 Haplogroup F(xK) 168

7.4.2.2 Haplogroup J2 169

7.4.2.3 Haplogroup O3 170

7.4.2.4 Haplogroup DE 172

7.4.2.5 Haplogroup R1a1 174

7.5 Summary 176

CHAPTER 8 DISCUSSION AND CONCLUSIONS 177

8.1 Introduction 178

8.2 The Han 178

8.3 Tibet and the Tibetans 181

8.4 Yunnan populations: the Miao, Yao and Kucong 184

8.4.1 The Miao 185

8.4.2 The Yao 186

8.4.3 The Kucong 187

8.5 The Silk Road and the arrival of Islam in China 190

8.6 The concept of ethnicity in China 194

8.7 Population genetics and population modelling 197

CHAPTER 9 RECOMMENDATIONS FOR FUTURE STUDIES 201

9.1 Introduction 202

9.2 Consequences and recommendations for future studies 202

REFERENCES 205

APPENDICES 226

Appendix A Primer Sequences 227

Appendix A.1 Autosomal STR PCR primer sequences 228

Appendix A.2 Y-chromosome STR PCR primer sequences 229

Appendix A.3 Y-chromosome SNP primer sequences 230

Appendix A.4 Mitochondrial HV-I PCR primer sequences 231

Appendix A.5 Primers for mtDNA 9bp indel detection in theCOII/tRNA^{LYS} ..... intergenic region 231

Appendix B Relevant raw data derived from the genomic analyses 232

Appendix B.1 mtdna HV1 sequences from the 9 study populations. 233

Appendix B.2 Y chromosome STR haplotypes 235

B.2.1 Bo'an Y-chromosome STR haplotypes 235

B.2.2 Dongxiang Y chromsome STR Haplotypes 236

B.2.3 Han Y-chromosome STR haplotypes 237

B.2.4 Hui Y-chromosome STR haplotypes 239

B.2.5 Kucong Y-chromosome STR haplotypes 240

B.2.6 Miao Y-chromosome STR haplotypes 241

B.2.7 Sala Y-chromosome STR haplotypes 242

B.2.8 Tibetan Y-chromosome STR haplotypes 243

B.2.9 Yao Y-chromosome STR haplotypes 243

Appendix B.3 Autosomal STR genotypes 241

B.3.1 Bo'an autosomal genotypes 244

B.3.2 Dongxiang autosomal genotypes 247

B.3.3 Han autosomal genotypes 250

B.3.4 Hui autosomal genotypes 255

B.3.5 Kucong autosomal genotypes 258

B.3.6 Miao autosomal genotypes 260

B.3.7 Sala autosomal genotypes 262

B.3.8 Tibetan autosomal genotypes 266

B.3.9 Yao autosomal genotypes 267

LIST OF FIGURES

Figure 2.1 The location of the three major East Asian Neolithic cultures 19

Figure 3.1 Neighbour-joining consensus trees for various European populations generated from autosomal STR data 35

Figure 3.2 3D PCA plot of Y-chromosome haplotype frequencies from selected populations in Asia, Melanesia and Australia (Kayser *et al.*, 2001a) 37

Figure 4.1 Map of China 57

Figure 4.2 Comparative linkage maps locating STR markers on chromosomes 13 and 15, with an average distance of ten centimorgans 68

Figure 4.3 Cytogenetic map of the Y-chromosome indicating the approximate location of the STR markers examined in this study 73

Figure 4.4 Partial genealogy showing haplogroup designations and the defining mutations of Y-chromosome SNPs 74

Figure 4.5 Simplified Eurasian mtDNA HV1 haplogroup network 83

Figure 5.1 Consensus unrooted Neighbour Joining Tree constructed from genetic distances calculated for autosomal microsatellite genotype data from the nine study populations 93

Figure 5.2 Multidimensional Scaling Analysis based on genetic distances calculated for autosomal microsatellite genotype data from the nine study populations 94

Figure 5.3 Unrooted Neighbour Joining Tree based on genetic distances calculated from Y-chromosome microsatellite haplotypes of the nine study populations 104

Figure 5.4 Multidimensional Scaling Analysis based on genetic distances calculated for Y-chromosome microsatellite haplotypes from the nine study populations 105

Figure 5.5 Unrooted Neighbour-Joining Tree based on mean Tamura-Nei genetic distances calculated for mtDNA HV1 sequence data from the nine study populations 112

Figure 5.6 Multidimensional Scaling Analysis scatter plot based on genetic distances calculated mtDNA HV1 sequence data from the nine study populations 113

Figure 6.1 PCA plot of the nine study populations based on mtDNA haplogroup frequencies 118

Figure 6.2 PCA plot of the nine study populations based on Y-chromosome haplogroup frequencies 128

Figure 6.3 Median Joining Networks of Y-chromosome STR haplotypes within selected haplogroups from the nine study populations 132

Figure 6.4 Multidimensional Scaling Analysis based on genetic distances calculated for Y-chromosome microsatellite data from the nine study populations 137

Figure 7.1 Estimated population structure from STRUCTURE analysis of autosomal STRs 155

Figure 7.2 Estimated intra-population structure from STRUCTURE analysis of autosomal STRs data in selected populations 157

Figure 7.3 Simplified schema of BATWING models utilised in the intra-haplogroup analyses 161

Figure 7.4 Consensus Bayesian phylogenetic population tree from the beta growth BATWING model 167

LIST OF EQUATIONS

Equation 3.1 Wright-Fisher Probability Sampling Distribution 24

Equation 3.2 Wright-Fisher Transition Probability 25

Equation 3.3 Expected Heterozygosity under WFN Conditions 25

Equation 3.3 Watterson's Estimate of θ 26

Equation 3.4 Wright's F statistics 27

Equation 3.5 Weir and Cockerham's F_{ST} 28

Equation 3.6 Observed Heterozygosity of an Individual in a Subpopulation 28

Equation 3.7 Expected Heterozygosity of an Individual 28

Equation 3.8 The Expected Heterozygosity of an Individual in an Equivalent Random Mating Population. 29

Equation 3.9 Nei's F statistics 29

Equation 3.10 Squared Euclidean Distance 30

Equation 3.11. Matrix of Euclidean Distances 31

Equation 3.12. Phi Statistics 32

Equation 3.13 Nei's Standard Genetic Distance 33

Equation 3.14 Goldstein's Delta Mu Squared Genetic Distance 34

Equation 3.15 Likelihood of a Phylogenetic Tree 39

Equation 3.16 Probability of Coalescence at Time t 42

Equation 3.17 Overall Probability of Coalescence 42

Equation 3.18 Probability of Coalescence Occurring in a Given Time Interval (Wilson et al., 2003) 43

Equation 3.19 Phylogenetic Likelihood 43

Equation 3.20 Discrete Coalescent Likelihood 43

Equation 3.21 Continuous Coalescent Likelihood 44

Equation 3.22 Expectation of a function of X 45

Equation 3.23 Approximation of the Expectation of a Function of X 45

Equation 3.24 Ergodic Average 46

Equation 3.25 Metropolis-Hastings Acceptance Probability 46

Equation 3.26 Likelihood Model of Beerli and Felsenstein (1999, 2001) 47

Equation 3.27 Likelihood Ratio Approximation 47

Equation 3.28 Metropolis-Hasting Acceptance Probability for Equation 3.27 47

Equation 3.29 Bayesian Theorem 48

Equation 3.30 Bayesian Metropolis-Hasting Acceptance Probability 49

Equation 3.31 Pritchard's Bayesian Clustering Posterior Probability 50

Equation 3.32 Likelihood of X (a), and the Prior of Z (b) 51
Equation 3.33 Approximation of the Prior of P 51
Equation 3.42 a-b Prior of Z for the Admixture Model 52
Equation 3.35 Prior of P Assuming Correlated Allele Frequencies 52
Equation 3.36 Prior of Z in the linkage model 53
Equation 5.1 Nei's Genetic Diversity. 86
Equation 5.2 Estimated Number of Alleles 99
Equation 5.3 Mean Nucleotide Diversity 106
Equation 5.4 Tajima's D statistic 108
Equation 7.1 Parameter Matrix for the MIGRATE-N Model 142
Equation 7.2 Likelihood of the Genealogy given the Parameters of the MIGRATE-N Model 142

LIST OF TABLES

Table 1.1 The official *minzu* of PR China 4

Table 4.1 Chinese dynasties 58

Table 4.2 Blood and DNA samples held in the Centre for Human Genetics, ECU 65

Table 4.3 PCR protocol for autosomal microsatellite markers 69

Table 4.4 PCR protocol for markers DYS19, DYS388, DYS392 and DYS393 75

Table 4.5 PCR protocol for markers DYS389, DYS390 and DYS391 75

Table 4.6 Y-chromosome biallelic PCR protocol 76

Table 4.7 mtDNA HV1 sequence amplification cycling conditions 79

Table 4.8 mtDNA HV1 sequence reaction thermal cycling conditions 80

Table 4.9 COII/tRNA^{LYS} intergenic region thermal cycling conditions 81

Table 5.1 Summary statistics from autosomal microsatellite analysis 87

Table 5.2 Inferred Group Structures for AMOVA analysis 90

Table 5.3 AMOVA analysis of autosomal genotypes 90

Table 5.4 Summary Statistics from Y-chromosome microsatellite analysis 96

Table 5.5 DYS388 allele frequencies in the nine study populations 98

Table 5.6 Summary of haplotype-sharing between the study populations 101

Table 5.7 AMOVA analysis of Y chromosome microsatellite diversity in the 9 study populations. 102

Table 5.8 Summary statistics from mtDNA HV1 sequence analysis 107

Table 5.9 AMOVA analysis of mtDNA HV1 sequence diversity in the nine study populations 111

Table 6.1 Table of mtDNA haplogroup frequencies in the nine study populations . 117

Table 6.2 Y-chromosome haplogroup frequencies in the nine study populations 123

Table 6.3 Rho statistics and estimated TMRCA of Haplogroups resolved from the nine study populations 134

Table 7.1 Matrix of Maximum Likelihood Estimates of Θ and M parameters for autosomal microsatellite data from the nine study populations 148

Table 7.2 Matrix of Maximum Likelihood Estimates of Θ and M parameters for mtDNA sequence data from the nine study populations 149

Table 7.3 Matrix of Maximum Likelihood Estimates of Θ and M parameters for Y-chromosome microsatellite data from the nine study populations 150

Table 7.4 Summary of inter-population STRUCTURE analysis 153

Table 7.5 Summary of intra-population STRUCTURE analysis 154

Table 7.6 Summary of marginal posterior distributions generated from BATWING inter-population analysis 164

Table 7.7 Summary of marginal posterior distributions produced by BATWING analysis of Kucong Haplogroup F(xK) samples. 168

Table 7.8 Summary of marginal posterior distributions produced by BATWING analysis of Han haplogroup J2 samples. 170

Table 7.9 Summary of marginal posterior distributions produced by BATWING beta growth model analysis of Han, Miao and Sala haplogroup O3 samples 171

Table 7.10 Summary of marginal posterior distributions produced by BATWING beta growth model analysis of Tibetan and Yao haplogroup DE samples 172

Table 7.11 Summary of marginal posterior distributions produced by BATWING standard coalescent model analysis of the Yao haplogroup DE samples . 174

Table 7.10 Summary of marginal posterior distributions produced by BATWING beta growth model analysis of Bo'an, Dongxiang and Sala haplogroup R1a1 samples 175

CHAPTER 1

INTRODUCTION

1.1 Introduction

The origin of *Homo sapiens* has long fascinated scientists. Since the publication of '*The Origin of the Species*' followed by '*The Descent of Man*' (Darwin 1859, 1871) there has been great interest in the origins and subsequent migrations of modern humans throughout the world. In recent times DNA-based techniques have revolutionised this search, which previously had been conducted largely by classical palaeontological and anthropological methodologies.

In 1991 the Human Genome Diversity Project (HGDP) was established to co-ordinate the collection and analysis of DNA samples from ethnic groups resident in different regions of the world (Cavalli-Sforza *et al.*, 1991). HGDP-related projects have been established in countries with extensive recorded histories, including China, Pakistan and Israel (Steinberg, 2000). While at times controversial, the collection of genetic data from the widely dispersed ethnic groups investigated in these sub-projects, especially in areas where the fossil specimens of modern humans have been found, has become vital in clarifying the origins and spread of the human species (Cavalli-Sforza *et al.*, 1994; Gebo *et al.*, 2000; Templeton, 2002).

The past and present size of the population forms an important aspect of this research in the study of genetic diversity in China. From past molecular genetic studies it was believed that modern humans (*Homo sapiens*) had migrated from Africa after the last Ice Age (60,000–100,000 bp), and either settled in southern East Asia or Central Asia before moving into China and Siberia (Nei and Roychoudhury, 1993; Cavalli-Sforza *et al.*, 1994; Cavalli-Sforza, 1998b; Chu *et al.*, 1998; Su *et al.*, 1999). An alternative model, referred to as the 'pincer' model, suggested that both routes played equally significant roles in the peopling of Asia (Ding *et al.*, 2000). A recent molecular study has proposed a more complex scenario, involving multiple migrations through different historical periods in history from both Southeast and Central Asia (Karafet *et al.*, 2001). Importantly, this study also concluded that future investigations into genetic diversity in China should consider historical, demographic and especially cultural factors in parallel with molecular data when assessing the genetic background of human colonisation of the country.

Historical evidence suggests that such an approach is prudent. Throughout recorded Chinese history, the number of residents of non-Han ancestry and their

political and commercial influence fluctuated greatly. However, in most periods ,non-Han communities were regarded as important contributors to the nation (Du and Yip 1993). Many of these communities had moved from their homelands, settled in China and intermarried with Han Chinese. As a result, over the generations they became accepted as fully-fledged Chinese citizens, even though they retained many of the customs and religious practices of their homeland and frequently their own language. There are now 56 officially recognised ethnic groups (*minzu*) within the PR China population of 1,300 million (Population Reference Bureau, 2004), with the majority Han population forming some 90% of the total (Table 1.1). The other 55 minority populations mostly live in peripheral and boundary regions of the country. In addition, there are ethnic groups, such as the Kucong (Yang, 1994), which are of more obscure origins and have not been accorded official governmental recognition.

Studies into genetic diversity among Chinese ethnic populations have mainly investigated factors such as selection, founder and/or bottleneck effects, genetic drift, and mutation, and addressed human migration on an evolutionary time-scale (Chu *et al.*, 1998; Su *et al.*, 1999; Ding *et al.*, 2000). To date, only limited attention has been paid to topics of anthropological and demographic interest, including endogamy, consanguinity, polygyny and polygamy, past and current population sizes, kin-structured migration and major population upheavals, all of which can shape the pattern of genetic diversity within the shorter time-frame of historical events (Karafet *et al.*, 2001; Zerjal *et al.*, 2001; Zerjal *et al.*, 2002). Yet in the past, demographic characteristics of this type have been responsible for important population movements and changes within China, including the quite recent and ongoing migration of Han Chinese into the border regions of Xinjiang, Yunnan and Tibet (Du and Yip, 1993).

An earlier study of autosomal and Y-chromosomal microsatellite diversity in the Han and the Hui Muslim community in Liaoning Province in northeast China (Black *et al.*, 2001) suggested that DNA polymorphisms could be profitably employed as genetic markers in the assessment of population factors, especially as indicators of the social and demographic histories of communities. The major finding of the study was that Y-chromosome and autosomal diversity differed significantly, with major Y-chromosome differences between the populations, whereas the autosomal data suggested more cohesive inter-population relationships.

Table 1.1 The official *minzu* of PR China

Ethnic Group	Population	Ethnic Group	Population
Achang	33,900	Manchu	10.67 million
Bai	1.86 million	Maonan	107,100
Blang	92,000	*Miao*	8.9 million
Bo'an	16,500	Mongba	8,900
Bouyei	2.97 million	Mongolian	4.76 million
Dai	1.6 million	Mulam	207,300
Daur	132,400	Naxi	308,800
De'ang	17,900	Nu	28,800
Dong	2.96 million	Oroqen	8,100
Dongxiang	513,800	Ozbek	12,400
Drung	7,500	Pumi	33,600
Ewenki	30,500	Qiang	306,100
Gaoshan	415,000	Russian	15,600
Gelo	579,000	*Sala*	104,500
Han	1,190 million	She	709,500
Hani	1.44 million	Shui	406,900
Hezhe	4,600	Tajik	41,100
Hui	9.8 million	Tartar	4,900
Jing	22,500	*Tibetan*	5.46 million
Jino	20,500	Tu	241,100
Jingpo	132,400	Tujia	8.03 million
Kazak	1.25 million	Uygur	8.39 million
Kirgiz	160,800	Va	396,600
Korean	1.9 million	Xibe	188,800
Lahu	453,700	*Yao*	2.64 million
Lhoba	2,900	Yi	7.76 million
Li	1.25 million	Yugur	13,700
Lisu	635,000	Zhuang	16.18 million

*Underlined italics indicte populations surveyed in the current study.
Data sourced from PR China 2000 Cbensus (National Bureau of Statistics of China, 2002).

It was concluded that the findings were best explained by inter-marriage of the male founders of the Hui community with Han Chinese females (Du and Yip, 1993; Lipman, 1997).

The aim of this project was to build upon and greatly extend the knowledge gained in the former study by the introduction and evaluation of new methodologies to further investigate the Hui and Han. At the same time a group of seven other widely divergent Chinese minorities was investigated. These populations include three additional Muslim populations resident in central China, the Bo'an, Dongxiang and Sala, and four southern populations that are of putatively more ancient origins the Kucong, Miao, Yao and Tibetans. The history, religion, language and marriage practices of these four southern Chinese populations contrast strongly with the Han and the Muslim study communities.

According to recent official estimates (National Bureau of Statistics of China 2002), the population sizes of the nine minorities are Hui 9.8 million, Bo'an 16,500, Dongxiang 513,800, Sala 104,500, Kucong unknown, Miao 8.9 million, Yao, 2.6 million and Tibetans 5.4 million, while the Han majority numbers at least 1190 million (Table 1.1).

1.2 Significance of the study

The study differs from previous molecular anthropological studies by taking into account socio-demographic effects and focusing on population characteristics, such as endogamy and historical gene flow, which could have disrupted the migration-drift equilibrium. In addition, detailed genomic analysis has not previously been undertaken on a number of the study populations. Methods and/or information from three genetic systems is provided to disentangle the genomic effects of evolutionary events, such as founder effects, bottlenecks and long distance dispersal, from the roles of historical and cultural influence.

Inclusion of the previously uninvestigated Kucong minority provided an excellent opportunity to evaluate the applicability of current genetic investigative techniques and theories in tracing ancestry and origins, where there is little available written historical and demographic evidence.

1.3 Aims of the Study

1. The generation of substantial new, population-specific genomic data on Chinese populations and its collation with results previously generated in the Centre for Human Genetics.
2. A comparison of these genomic data with information from historical, demographic and anthropological sources, to gauge the overall structure of each study population.
3. The identification of migration patterns, patrilocality and population continuity in each of the minority communities, with the patterns and intensity of paternal gene flow examined by reference to Y-chromosome (paternal lineage), and maternal gene flow by reference to mtDNA (maternal lineage) loci.
4. An investigation into the influence of socio-demographic factors on the genetic structure of human populations, and consequently on the ability to trace ancestry.
5. An investigation into the applicability of Coalescent based methodologies as applied to genetic history, and including socio-demographic effects.

1.4 Research Issues and Questions

The study explores the following issues with regard to the broad field of human population genetics, as well as answering questions specifically focused on the nine study populations. The questions addressed in the study are :

1. To what extent do the bi- and uniparental DNA pools of the each community reflect their cultural and genetic characteristics, and does the genetic information indicate any common heritage or historical admixture?
2. Is it possible to discern sex-related differences in migration patterns, patrilocal and/or matrilocal contributions, and population continuity in each of the minority communities?
3. What depth of population ancestry can be determined in the different minority populations?
4. What are the effects on genetic variation of events that occurred within a historical time-frame, as opposed to across an evolutionary time scale?
5. How applicable are current 'classical' methodologies in modelling the complex population structures presented in this study, e.g., (the effect of factors such as endogamy and multiple migration)?
6. Can newly developed Likelihood and Bayesian methods be used to infer genetic factors such as sex-mediated gene flow, inbreeding, drift, and migration from data resolved from these complex population structures?
7. What, if any, new insights will the study provide in terms of the human history of China?

CHAPTER 2

REVIEW OF MOLECULAR ANTHROPOLOGY

2.1 Molecular anthropology and the search for human origins

Molecular anthropology is focused on human demographic history and is a relatively new field. However, the study of genetic diversity to examine human evolutionary history, mainly tracing human origins over continental areas, is older and better established.

The first survey of global human genetic diversity was the pioneering work by Cavalli-Sforza and Bodmer, (1971). In a subsequent volume, (Cavalli-Sforza *et al.*, 1994), a more comprehensive survey of human genetic diversity around the globe was mapped and used to determine the origins of modern humans and their migrations. Studies of this nature incorporate palaeontology, archaeology, history and geography with anthropomorphic and genetic data, including the distribution of blood groups and protein polymorphisms.

As a result of this ongoing research two major theories for the origin of *Homo sapiens* have been proposed. These are the Out of Africa Theory (OOA) and the Multiregional Evolution Hypothesis (MEH). The OOA theory states that *Homo sapiens* evolved in Africa and then migrated out of Africa some 50,000 to 200,000 years before present, completely replacing earlier hominid species such as *Homo erectus* (Cann, 2001). In contrast, the MEH argues that *Homo sapiens* evolved directly from archaic hominid species in several different locations. Gene flow among these populations and the action of natural selection for advantageous genes has thus been responsible for genetic homogeneity of the species.

The investigation of human diversity has been greatly advanced by the development of DNA markers, with the development of polymerase chain reaction (PCR) and sequencing technologies allowing high-throughput, large-scale analysis of the human genome. An early application of these technologies was the study of human mitochondrial genome diversity by Cann *et al.* (1987) who concluded that all modern humans share a female ancestor who lived in Africa roughly 200,000 YBP. The haplotype trees generated in this study were found to be flawed (Templeton, 1992), but subsequent studies have generated more reliable haplotype trees which have confirmed the hypothesis of an African origin for modern humans (Torroni *et al.*, 1992; Chen *et al.*, 1995). In addition, recent Y-chromosomal studies have also

reached this conclusion (Ruiz Linares *et al.*, 1996; Su *et al.*, 1999; Underhill *et al.*, 2000; Underhill *et al.*, 2001).

Furthermore, many of these studies infer an effective population size for the human species of around 10,000. This figure is indicative of a small ancestral base and hence strongly suggestive of a single recent origin. By comparison, the MEH requires the existence of hundreds of thousands of ancestors for most of the last million years (Harpending *et al.*, 1998). Without this large initial population, gene flow among the populations assumed to have been distributed widely over the temperate and tropical Old World would have been impossible. Therefore an effective population size of 10,000 would seem to discredit a multiregional model for human origins.

A more complicated model has been proposed which could in part accommodate aspects of the multiregional hypothesis within an Out of Africa framework (Templeton, 2002). The model was based on the analysis of mtDNA, Y-chromosomes and some autosomal markers, and it includes multiple migrations out of Africa and subsequent breeding between the different migrant populations. It stresses the importance of gene flow, admixture, changing selection pressures and resulting directional morphological change over direct replacement.

Concerns have been raised about the conclusions drawn (Satta and Takahata, 2002), and it has been argued that the nested clade analysis used may not be quantitatively reliable. Furthermore, it was suggested that while multiple migrations may well have occurred, the model formulated by Templeton (2002) could also be interpreted as showing that the most recent migration could have replaced previous migrations, as stipulated in the OOA theory.

These recent arguments about modern human origins show that, given the ability to define and assess fine resolution genomic markers and the development of greater computing power, resolution of the origins of *Homo sapiens* becomes more complex. In fact, most genetic anthropology studies now do not explicitly stipulate any specific model for the origins of human populations (Relethford, 2001). To further explore our origins it is increasingly evident that population structure plays a significant role in the study of our evolving past (Goldstein and Chikhi, 2002). Thus to resolve genetic signals pertaining to our origins, it is prudent to explore the effects

on the modern human gene pool of more recent historical processes and demographic changes.

2.2 Molecular systems in molecular anthropology

Exploration of demographic effects on genetic structure involves the knowledge of the properties of different genetic markers from different genetic systems. The major DNA markers used in molecular anthropology can be sub-divided into four distinct classes - autosomal, X-chromosome, Y-chromosome and mtDNA markers. Various types of markers are used, including Restriction Fragment Length Polymorphisms (RFLPs), Short Tandem Repeats (STRs) or microsatellites, Single Nucleotide Polymorphisms (SNPs), and specific DNA marker anomalies, e.g., the presence of the 355bp Y-chromosome Alu sequences (YAP), and the deletion of specific sequences such as the 9bp deletion in the COII/ $tRNA^{Lys}$ mtDNA region.

Each class of DNA marker uncovers a different aspect of human genetic diversity, with mtDNA markers being used to trace maternal ancestry and Y-chromosome markers applied to the study of paternal lineages. The following discussion will be divided into two sections dealing separately with mtDNA and Y-chromosome diversity, as these marker systems will be the main focus of the proposed study.

2.2.1 The human mitochondrial genome

The mitochondrial genome consists of 16,569 base pairs (bp) of highly conserved and organised DNA sequence. The genome replicates autonomously and is inherited exclusively from the mother. The mtDNA genome was the first portion of the Human Genome to be completely sequenced (Anderson *et al.*, 1981).

Mitochondrial DNA (mtDNA) has been widely used to study genetic relationships among closely related populations because of its rapid rate of evolution and maternal mode of inheritance (Horai *et al.*, 1993; Horai *et al.*, 1996). In recent mtDNA studies, it has been common to directly sequence the non-coding (D-Loop) region of the mtDNA, which has apparently evolved several times faster than other parts of the mtDNA genome (Vigilant *et al.*, 1989; Horai *et al.*, 1996). Sequence analysis of this region therefore offers high resolution that discriminates between very closely related individuals. Subsequent studies of mtDNA diversity used

restriction fragment length polymorphisms (RFLPs) to survey the entire mitochondrial genome and investigate relationships between continental populations, in an effort to reconstruct human female evolutionary history (Richards *et al.*, 1998; Herrnstadt *et al.*, 2002).

In addition, sequencing of the Hypervariable Segment 1 (HVS1) of the D-loop that rapidly accumulates mutations has been employed in a number of studies. It has been shown that most RFLP haplogroups can be recognised by specific nucleotide motifs in the control region (Torroni *et al.*, 1996; Richards *et al.*, 1998; Macaulay *et al.*, 1999). After over 20 years of investigation, there are now data on whole mtDNA genomes and various mtDNA regions available for many human populations. This data is stored on major databases such as Genbank (http://www.ncbi.nlm.nih.gov/Genbank/index.html).

The haploid maternal inheritance of mtDNA means that it is particularly sensitive to reductions and expansions in population size, in comparison to the autosomal chromosomes. Moreover, the high mutation rate and apparent lack of selection in the D-loop region result in the rapid differentiation of maternal lineages between populations. Thus, the investigation of sequences within a population and between closely related populations could provide meaningful information on demographic and social history.

An example of demographically meaningful mtDNA data is provided by the 9bp deletion caused by the loss of one copy of the 9-bp tandem repeat sequence (CCCCCTCTA) in the COII/tRNA^{Lys} intergenic region of human mtDNA. Originally thought to be Asian-specific, the deletion has been found at varying frequencies in Asian, Pacific, Amerindian, African and European populations (Yao *et al.*, 2000). However, the 9-bp deletion has an increasing 'west to east' clinal geographic frequency distribution, making it more common in East Asian than in European populations. Thus a recent study concluded that the 9-bp deletion is useful as a population-specific marker for tracing the migration patterns of Chinese populations, regardless of its low frequency occurrence in other global populations (Yao *et al.*, 2000). The 9-bp deletion has been included as a marker for one of the haplogroups of the RFLP-HV1 haplogrouping systems, Haplogroup B, found predominantly in Amerindian and Asian populations. It is also recurrent and is present on other Haplogroups, though not in East Asia. Therefore a survey of the

frequency of the deletion in Chinese ethnic populations may help to prove or disprove the Han female ancestry that historians and others have claimed for many of these populations (Du and Yip, 1993; Gladney, 1996; Lipman, 1997).

2.2.2 Genetic diversity of the human Y-chromosome

The use of Y-chromosome haplotypes in investigating genetic diversity and tracing ancestry is more recent than that of mtDNA, as most Y-chromosome markers were discovered subsequent to developments in genomic screening technologies.

By their nature, Y-chromosomes like the mtDNA genome are effectively haploid, with only a small portion of the chromosome that can undergo recombination with the X-chromosome. The lack of recombination means that genetic diversity is more limited in Y-chromosomes than in the autosomes. In turn, the absence of recombination, and the fact that the effective number of Y-chromosomes is one quarter the number of autosomal chromosomes, increases the effect of genetic drift. These properties of Y-chromosomes have proved very useful in elucidating genetic differences between closely related populations whose times of divergence have been relatively short (de Knijff *et al.*, 1997), and in tracing human origins (Underhill *et al.*, 2001).

The identification of different Y-chromosome markers has allowed geneticists to provide a unique and detailed insight into human genetic variation via the construction and analysis of Y-chromosome haplotypes (Cooper *et al.*, 1996; Roewer *et al.*, 1996; Underhill *et al.*, 2001; Wells *et al.*, 2001; Cruciani *et al.*, 2002; Lell *et al.*, 2002). Examples of the applications of Y-chromosome analysis include investigation of the origins of Amerindians and their relationship with modern Siberian populations (Lell *et al.*, 2002), the genetic history of gene flow between north-western Africa and the Iberian peninsula (Bosch *et al.*, 2001), the genetic origins of the Roma (Gresham *et al.*, 2001), and the origins of modern humans as a species (Underhill *et al.*, 2001).

The first Y-chromosome polymorphisms were reported nearly two decades ago (Casanova *et al.*, 1985). Progress in discovering further examples was initially slow because conventional DNA polymorphisms have been difficult to identify on the Y-chromosome, and those that have been discovered often have proved to be of limited informativity (Jobling and Tyler-Smith, 1995). However, a series of

polymorphic microsatellites have been developed and tested on many different population samples from around the world (de Knijff *et al.*, 1997; Kayser *et al.*, 1997). These microsatellites (DYS19, DYS388, DYS389I+II, DYS390, DYS391, DYS392 and DYS393) are highly polymorphic, and they have proved appropriate as tools both in forensic studies and for the analysis of Y-chromosome origins (Kayser *et al.* 1997).

More recent investigations have focused on biallelic markers, especially Single Nucleotide Polymorphisms. SNPs were first described as RFLPs two decades ago, but the development of high-throughput analysis and the consequent generation of SNP data is a more recent phenomenon (Brookes, 1999). The technologies developed include denaturing high performance liquid chromatography (Underhill *et al.*, 1997) and matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry (Griffin and Smith, 2000). These methods allow SNPs to be typed efficiently and relatively cheaply, with the construction of increasingly dense genomic maps. By early 2005, over 10 million SNPs from all 23 chromosomes, including around 40,000 mapped on the Y-chromosome, had been submitted to the dbSNP database at the National Center of Biotechnology Information (U.S.A).

SNPs can be detailed as a single base pair position in genomic DNA at which different sequence alternatives (alleles) exist, with the least abundant allele present at a frequency in the population of 1% or greater. In principle, SNPs can be bi-, tri- or tetra-allelic. However tri- and tetra-allelic variants are rare, and so biallelic SNPs are usually the rule (Brookes, 1999).

From a molecular anthropological standpoint, SNPs and related biallelic markers have proved extremely useful. For example, a set of 280 Y-chromosome biallelic markers were classified into continental Y-chromosome haplotypes numbered I – X (Underhill *et al.*, 2001). These haplotypes have been shown to exist at varying frequencies among different broad continental groupings, e.g., Asian, African, European and Amerindian populations (Underhill *et al.*, 2001). While predominately characterised by SNPs, the haplogrouping system also uses other biallelic makers. One such marker is M1, better known as YAP (Y-chromosome Alu Polymorphism), a Y-chromosome Alu insertion element 355bp in length. The presence of this element defines haplogroups that predominate in Asian populations

(Horai *et al.*, 1996) and is also found in some North African populations (Arredi *et al.* 2004).

A more comprehensive haplogrouping schema has recently been developed. The Y Chromosome Consortium (http://ycc.biosci.arizona.edu/), a group involved in a collaborative effort to study genetic variation on the human Y-chromosome, has developed a binary marker haplogroup system from several unrelated and non-systematic biallelic haplogroup nomenclatures (Y Chromosome Consortium, 2002). The proposed haplogroup system consists of a fusion of the systems developed by Underhill *et al.* (2001), Jobling and Tyler-Smith (2000), Hammer *et al.* (2001), Semino *et al.* (2000) and Karafet *et al.* (2001), and also a system developed and refined by Su *et al.* (1999) and Capelli *et al.* (2001). The new haplogrouping system gives researchers a universal Y-chromosome nomenclature that should allow the direct comparison of data from different studies, in a similar manner to mtDNA studies.

The development of Y-chromosome haplogrouping has led to a narrowing of focus for molecular anthropology from a global view to more localised studies. One example is research into the Jewish Y-chromosome rabbinical lineages (Thomas *et al.*, 1998; Behar *et al.*, 2003). The diversity of haplotypes consisting of six microsatellite markers and 6 biallelic markers were investigated in various Jewish populations and the rabbinical families within them. Of 112 haplotypes found, one was strikingly frequent in both Ashkenazi and Sephardic Cohanim rabbinical groups and thus named the Cohen Modal Haplotype. Such a high frequency of the one haplotype in one rabbinical group in two disparate Jewish populations suggested a shared common origin of these rabbinical groups derived from a common ancestral chromosome. Equivalent restricted diversity and origins were also found in a later study on another rabbinical group, the Levites (Behar *et al.*, 2003).

A similar pattern of restricted male ancestries are also demonstrated in East Asia, with the finding of a tight cluster of Y-chromosome haplotypes at high frequencies throughout a large region of Asia from the Pacific to the Caspian Sea. It was concluded that this haplotype probably represents a lineage carried by male-line descendants of Genghis Khan and his male relatives, preserved by a novel form of social selection based on the prestige of claiming descent from the 'Great Khan' (Zerjal *et al.*, 2003).

The mixture of global and regional-based studies demonstrates the flexibility and power of Y-chromosome markers in defining different aspects of human history and population structure, as well as highlighting a more complex pattern of ancestry than previously recognised.

2.2.3 Autosomal and X-Chromosome Approaches

While the majority of molecular anthropology studies have been based on the haploid mtDNA and Y-chromosome systems, there has been growing focus on the X-chromosome and autosomal systems. However, many of these studies have not achieved either the population depth or worldwide breadth of mtDNA or Y-chromosome studies.

The use of autosomal microsatellites in global studies similar to those completed in Y-chromosome and mtDNA studies, have shown results in agreement with the OOA hypothesis, with high diversity in African populations and a decline in diversity proportional to increasing distance from Africa (Rosenberg *et al.*, 2002; Ayub *et al.*, 2003). Although general global autosomal microsatellite studies are in broad agreement with OOA predictions, the investigation of specific autosomal systems show different results. One example is provided by an exploration of variation within the β-globin gene region on chromosome 11. A survey of the sequence diversity of this gene region in various populations showed a distinct pattern, with the conclusion that there were two distinct archaic origins for Modern Humans, not one. These origins were in Africa with a MRCA dating to around 800,000 YBP (Harding *et al.*, 1997). The result is in contrast to the widely accepted OOA concept of an initial migration of *Homo sapiens* around 100,000 years ago. Instead, the β-globin gene data suggest an earlier dispersal of early *Homo sapiens* out of Africa and the subsequent inclusion of non-African pre-*sapiens* population groups into the gene pool.

The β-globin gene region is well known for the functional variants that occur, including the thalassaemias and other haemoglobinopathies, with elevated frequencies of specific alleles where *Plasmodium falciparum* is endemic because of the selective advantage conferred on heterozygous genotypes. Thus it has been argued that the findings of the study by Harding *et al.* (1997) are affected by selection, and so may be unreliable for calculating MRCA estimates. However, such

selection was not evident in the diversity study as, it was argued, the selective processes had only occurred in the last 10,000 years in concert with the development of *falciparum* malaria. Under such circumstances any effect from selection could be subsumed by 800,00 years of mutation and genetic drift (Harding *et al.*, 1997).

Further evidence of differing MRCA estimates from OOA expectations was shown in an X-chromosome study, which investigated the global diversity of 35 polymorphic sites and one microsatellite situated in an 8kb region of the dystrophin gene *dys44* (Labuda *et al.* 2000; Zietkiewicz *et al.* 2003). Three distinct lineages were identified, including one in Africa which was subsequently found in non-African populations, and suggested a migration out of Africa from 27,000 to 56,000 YBP. A second lineage was exclusively African and a third lineage was absent in Africa, yet dated older than the first lineage at around 160,000 years. From the results it was concluded that this genetic system represented a 'mosaic' of different contributions to modern human genetic diversity, effectively displaying different stages of ancient human evolution and migration.

For these reasons, it has been proposed that in the investigation of human origins, multiple systems should be examined, each of which illustrates different stages of human history. It also has been concluded by many researchers that a focus on intra-population substructure, demography and historical events is important in constructing a comprehensive model of modern human origins (Cann, 2001; Goldstein and Chikhi, 2002; Zerjal *et al.*, 2002).

2.3 Molecular anthropology and PR China

China is a prime example of the importance of population structure in molecular anthropology and has become a major focus in the field, due in part to the fact that the country is home to approximately 20% of the world population (Cavalli-Sforza, 1998a). Studies into East Asian genetic diversity were initially directed at the route(s) of the earliest dispersals of anatomically modern humans into China, and their hypothesised subsequent migration to Japan, Siberia and the Americas (Karafet *et al.*, 2001).

The early molecular genetic studies into Chinese diversity similarly concentrated on the first migrations into East Asia. In concert with archaeological evidence, it is hypothesised that modern man first arrived in China via Southeast

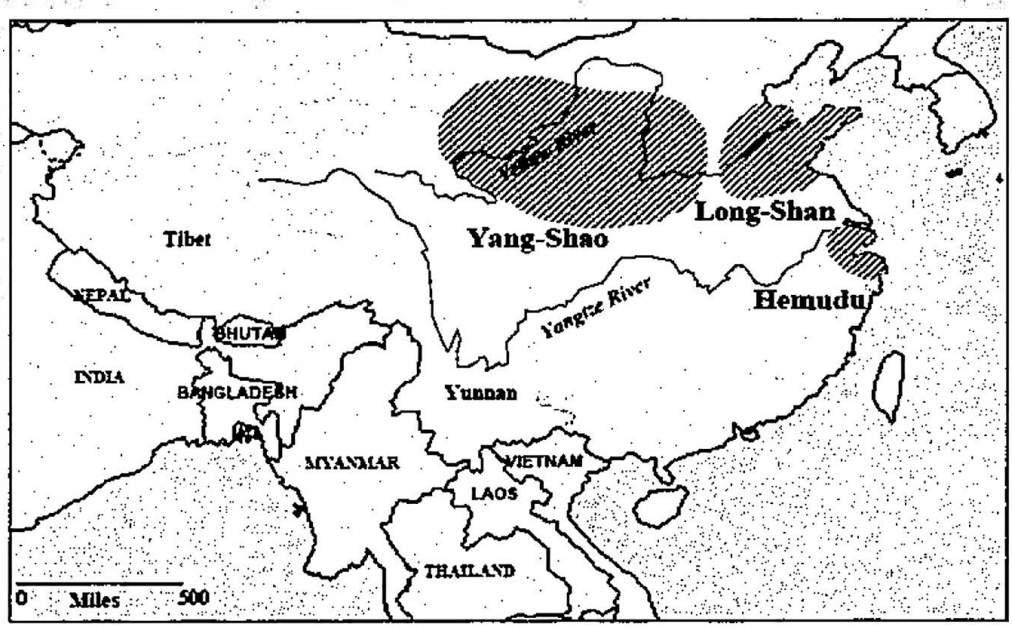
Asia at approximately 50,000 – 40,000 YBP (Su *et al.*, 1999; Su *et al.*, 2000; Underhill *et al.*, 2001; Deng *et al.*, 2004). In fact, it was also suggested that a possible second wave of migrations via the north occurred some 10,000-20,000 years later (Underhill *et al.*, 2000; Underhill *et al.*, 2001; Underhill, 2004).

In tracing the migration of *Homo sapiens* into and within East Asia a number of studies have produced contradictory theories in East Asian origins. Chu *et al.* (1998) and Su *et al.* (1999) suggested a major north-south division in Chinese populations, with the northern populations originating in the south of the country. However, Ding *et al.* (2000) found no molecular evidence of separate internal groupings and claimed that later gene flows from Central Asia had exerted a major effect on Chinese populations.

Another study indicated that the origins of East Asian populations might be more complicated than suggested by these earlier studies (Karafet *et al.*, 2001). They hypothesised a multidirectional pattern of migration and settlement, with both south to north migration and a major migration from Central Asia, primarily to northern China. With the accumulation of data, this scenario allowing for multiple ancient migrations into and within China seems to be the most likely (Underhill *et al.*, 2001; Deng *et al.*, 2004; Wen *et al.*, 2004a; Wen *et al.*, 2004b).

As a result of these various studies, it has become clear that the major factor which had the greatest effect on current Chinese genetic diversity is a more recent event, the Neolithic expansion which occurred ~10,000 YBP. The Neolithic period in East Asia represents a major cultural change, which archaeological evidence shows began in the upper and middle regions of the Yellow River (Cavalli-Sforza *et al.*, 1994; Su *et al.*, 1999; Su *et al.*, 2000). This event was the shift from a Palaeolithic hunter-gather lifestyle to the development of millet and rice agriculture, and the resultant formation of permanent settlements. The Yang-Shao, Hemudu and Long-Shan cultures are believed to be the three earliest examples of East Asian Neolithic cultures, based on their large archaeological deposits (Figure 2.1). It is from these cultures that people are believed to have migrated throughout China and even Southeast Asia, accompanied by their knowledge of agricultural technology.

Figure 2.1 The Location of the three major East Asian Neolithic cultures



(adapted from Su *et al.* 2000)

From a genetics perspective, the gene pool of these Neolithic migrants became predominant in East Asia, thus resulting in other ancestral lineages either becoming rare or extinct, a pattern most clearly demonstrated in Y-chromosome studies (Su *et al.*, 2000; Deng *et al.*, 2004). It can also be shown that population movement and interaction after the initial Neolithic expansion event has also had an effect on the formation of present-day East Asian genetic diversity. For example, this conclusion was reached by Yao et al. (2002a) in a study of mtDNA genomic diversity among 10 minority ethnic populations in Southwest China. A more recent study by Wen *et al.* (2004a) which utilised both mtDNA and Y-chromosome data concluded that more recent historical migrations occurring from the Western Jin Dynasty (245-316 AD) to the southern Song Dynasty (1127-1278 AD) had a significant effect on current genetic diversity of the Han. Studies have also suggested more recent historical immigration from regions such as Central Asia, Mongolia and the Middle East, and its effect on the structure of the current East Asian gene pool (Karafet *et al.*, 1999; Black *et al.*, 2001; Karafet *et al.*, 2001; Wang *et al.*, 2003).

It can be concluded that in further investigating the molecular anthropology of East Asia, more complete models should be considered that include both paternal and maternal ancestry, coupled with information on major historical population migrations and the more ancient events inferred from archaeological investigations.

CHAPTER 3

REVIEW OF POPULATION GENETIC THEORY

3.1 Introduction

The field of molecular anthropology involves the integration of genomic analysis with the development and application of population genetic models in the investigation of modern human populations. In modern population genetics practice numerous models representing different aspects of genetic systems are available, including population growth, migration and population structure. As computational power has increased, more complex and intensive modelling has become possible, with coalescent theory occupying a position of central importance. In this chapter a broad overview is provided of the various population genetic models and techniques employed in the present study, and for molecular anthropology in general.

3.2 The Neutral Theory

Assumption of the selective neutrality of mutations is central to all models and techniques presented in the chapter. That is, the assumption that in micro-evolutionary studies, such as those in molecular anthropology, one can in practice ignore the effects of selection. Before the development of molecular population genetics it was believed that genetic variation was maintained in equilibrium by balancing selection. With the development of protein electrophoresis and the subsequent elucidation of the structure of DNA, many researchers concluded that significantly more genetic variation existed than could be reasonably accounted for by balancing selection alone, leading to what is now called the neutralist-selectionist debate. It was Kimura (1968, 1983), and King and Jukes (1969) who derived the alternative neutralist theory, which posited that the majority of changes at the DNA level were of little or no functional consequence to the organism, therefore a large majority of the mutations observed in the genome were selectively neutral. The main features of the theory can be summarised as:

1. The majority of mutations are either strongly deleterious, or of no selective importance. Deleterious mutations are rapidly removed from the population, so most variation within species and differences between species are the result of neutral mutations;

2. The rate of molecular evolution due to neutral mutations is identical to the neutral mutation rate; $k=f_{neutral}\mu$, where k is the rate of substitution, $f_{neutral}$ is the proportion of all mutations that are neutral, and μ is the mutation rate;
3. The level of polymorphism in a population is a function of the neutral mutation rate and the effective population size;
4. Polymorphisms are transient, i.e., they are on their way to loss or fixation, rather than being balanced by selection.

Since the neutral theory was proposed some refinements have been introduced as more genetic markers have been introduced. The most notable is the 'nearly' neutral model that allows for the involvement of selection in certain situations (Ohta, 1995). However, the general consensus amongst population geneticists remains that much of the variation at the DNA level is the result of effectively neutral mutations.

The practical importance of the neutral theory of molecular evolution has been confirmed with the development of investigative systems such as Restriction Fragment Length Polymorphisms, the non-coding region of the mtDNA genome, Short Tandem Repeat (STR) markers and Single Nucleotide Polymorphisms (SNPs). The assumption of neutrality has allowed researchers to map genes using these neutral markers to establish genealogical models of human populations, and thus explore the genealogical process back to ancient ancestors without needing to account for the complex effects of selection.

3.3 The Wright-Fisher Neutral Model

The basic null model for reproduction and genetic drift in population genetics is the Wright-Fisher Neutral (WFN) model of random genetic drift which was developed implicitly by Fisher (1930) and explicitly by Wright (1931), and further refined to its modern form after the establishment of Neutral Theory.

In essence the model includes a number of assumptions from Hardy-Weinberg equilibrium (HWE) theory, with the notable difference that the WFN model assumes finite population size. As with Hardy-Weinberg, almost none of the

following WFN assumptions hold true for any single biological, let alone human, population:

1. Constant diploid population of size *N (2N alleles)*;
2. Non-overlapping generations;
3. Random mating with respect to both geographical location and genotype;
4. Sexual reproduction with the possibility of selfing;
5. No migration to or from other populations;
6. Mutations are neutral and occur at a constant rate, μ, per generation.

In such a theoretical population, reproduction works by randomly selecting a gamete from one individual, and then randomly selecting a gamete from a second individual (which may be the same, or differ from the first) to produce one offspring. This process is repeated until a daughter population of individuals has been generated. Mathematically, the basic assumptions of the model are as follows. Say two alleles are randomly chosen from the population of interest, named A_1 and A_2; we then define N as the number of gametes (in a diploid population the size is $2N$ gametes), X_t as the number of A_2 alleles at time t and $p_{ij}=P[X_{t+1} = j \mid X_t=i]$. As previously indicated, it is assumed in the WFN model that gametes are chosen randomly each generation from an effectively random gene pool that reflects parental allele frequencies. Therefore the sampling can be considered binomial and so:

$$p_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(1-\frac{i}{2N}\right)^{2N-j}$$

Equation 3.1 Wright-Fisher Probability Sampling Distribution

where the allele frequency of allele A_1 is $i/2N$ and the allele frequency of A_2 is $1- i/2N$. The equation above is more commonly utilised as a transition probability of a Markov chain notated as:

$$T_{ij} = \binom{2N}{j} p^j q^{2N-j}$$

Equation 3.2 Wright-Fisher Transition Probability

The transition probabilities are then put in a square matrix, T, with elements T_{ij} giving the transition probability from state i to state j for $i, j=0,1,2,...,2N$. Allowing multiple alleles would then involve constructing a n-dimensional matrix. This is a Markov chain which allows exploration of the model into multiple generations via iteration of the chain, and giving the expected outcome of the pure drift process (Hartl and Clark, 1997; Hedrick, 2000).

3.3.1 Summary Statistics under the WFN model

The WFN model only considers the effect of genetic drift on allele frequency in a population. However, this basic model possesses convenient properties that enable population geneticists to investigate other important factors. An example of Wright-Fisher based genetic inference is estimation of the fundamental population-genetic parameter θ. This parameter is a compound parameter as, assuming WFN conditions, $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate. The importance of estimating θ is its use as a estimate of mutation rate or effective population size. Assuming a WFN model, one can estimate θ as follows:

$$H_e = \frac{1}{1+\theta}$$

Equation 3.3 Expected Heterozygosity under WFN Conditions

where H_e is the expected heterozygosity of a population at equilibrium between drift and mutation. Using particular properties of various molecular markers, it is possible to generate an estimate of θ directly from molecular data.

One example is Watterson's (1975) estimate of θ based on the number of segregating sites from a sample of non-recombining DNA sequences, which is defined as:

$$\hat{\theta}_S = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

Equation 3.3 Watterson's Estimate of θ

Other summary statistics have been developed including π, the average pair-wise sequence difference. The calculation of summary statistics emanating from a WFN null model now predominates in molecular anthropology studies. Their use has allowed researchers to define and test the effects of variables such as inbreeding, population amalgamation, migration and varying mutation rate, on the genetic diversity of populations, with the backing of rigorous statistical analysis.

3.3.2 Population differentiation

An important set of summary statistics in molecular anthropological studies is that developed for assessing molecular evidence for population differentiation. When a population is divided into isolated subpopulations, heterozygosity is lower than if the population was undivided. Founder effects acting on different populations generally lead to subpopulations with allele frequencies that differ from those of the larger population. Therefore, since the allele frequency in each generation represents a sample of the previous generation's allele frequency, there will be greater sampling error in these small groups than in a larger undifferentiated population. Hence, genetic drift will push these demes toward different allele frequencies and to allele fixation more quickly than would occur in a larger undifferentiated population.

The decline in heterozygosity due to subdivision within a population is traditionally quantified using Wright's F statistics. F statistics are a measure of the difference between the mean heterozygosity among sub-populations, and the potential frequency of heterozygotes if all members of the population mixed freely and non-assortatively (Hartl and Clark, 1997).

3.3.2.1 Wright's F_{ST} model

These F statistics, or more correctly, Fixation indices, were first defined by (Wright, 1951). His approach was based on correlations between two uniting gametes. It assumed random sampling and equally related populations in which three F statistics were defined (were in effect correlation coefficients), and used to allocate genetic variability to the total population level (T), subdivisions of that population (S), and individuals (I). The resulting three measures were termed F_{IT}, F_{ST} and F_{IS} and their relationship is as follows:

$$(1-F_{IT}) = (1-F_{IS})(1-F_{ST})$$

Equation 3.4 Wright's F statistics

F_{ST} is a measure of genetic differentiation over subpopulations and is always positive. F_{IS} and F_{IT} are measures of the deviation from Hardy-Weinberg proportions within subpopulations, and in the total population respectively. F_{IS} values ranges between -1 and +1 indicating all heterozygotes and all homozygotes respectively. The F_{IT} statistic is rarely used. These statistics are usually calculated for different genetic markers, and then averaged across all loci, and all populations.

While Wright developed the model, there has been much argument about how to calculate the different indices. The three most common methods in human population genetics are discussed below.

3.3.2.2 Weir and Cockerham's approach

An extension of Wright's original statistics which takes into account multiallelic, multilocus data is by the partioning of variance, known as ANOVA. (Weir and Cockerham, 1984). This F_{ST} crucially still assumes random sampling and equally related populations.

The statistic is described as follows:

$$F_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}$$

Equation 3.5 Weir and Cockerham's F_{ST}

where σ_a^2 is the component of variance among individuals within populations, σ_b^2 is the component of variance among populations within groups, and σ_c^2 is the component of variance among groups within the total population, and where the total variance is given by $\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$.

Confusingly, while Weir and Cockerham refer to their statistic as θ, the coancestry coefficient, most researchers continue to simply use the term F_{ST}.

3.3.2.3 Nei's G_{ST}

Most natural populations have some phylogenetic relationship dependant upon population structure. Therefore such populations are not random samples from a set of many equally related populations. For this situation, an alternative method for multiallelic markers was developed (Nei, 1973; Nei and Roychoudhury, 1973; Nei, 1977). This method is suitable for any situation, regardless of phylogenetic relationships.

Nei (1973; 1977), defined statistics for multiallelic information from any number of subpopulations as follows:

$$H_I = \sum_{i=1}^{k} H_i / k$$

Equation 3.6 Observed Heterozygosity of an Individual in a Subpopulation

where H_i is the observed heterozygosity in subpopulation i from k subpopulations;

$$H_S = 1 - \sum_{i=1}^{k} \bar{p}_{i,s}^2$$

Equation 3.7 Expected Heterozygosity of an Individual

where $p_{i,s}$ is the frequency of the ith allele in subpopulation s, and

$$H_T = 1 - \sum_{i=1}^{k} \bar{p}_i^2$$

Equation 3.8 The Expected Heterozygosity of an Individual in an Equivalent Random Mating Population.

where p_i is the frequency of allele i averaged over k subpopulations. The inbreeding coefficients are then variously defined as:

$$F_{IS} = \frac{\overline{H}_S - H_I}{\overline{H}_S}$$

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

Equation 3.9 Nei's F statistics

The F_{ST} in this model is more commonly defined as the coefficient of gene differentiation, or G_{ST}.

Nei further refined his models to be independent of the ploidy of the organism. In summary, he disregarded genotypic frequencies and considered the genetic decomposition of genomic variation into inter-population and intra-population variation. Therefore, instead of dealing with heterozygosity, we are dealing with what Nei termed gene diversity (Nei, 1987). It is the G_{ST} calculated using gene diversities to which most researchers refer.

3.3.2.4 Analysis of molecular variance

These previous approaches to population subdivision assume Mendelian gene frequencies and do not involve mutational differences between different markers or the fact that different types of marker display different mutational behaviours. These properties are taken into account by the method termed Analysis of Molecular Variance (AMOVA) (Excoffier *et al.*, 1992).

AMOVA is a method of estimating population differentiation directly from molecular data and testing hypotheses based on this differentiation. In part, it was developed from earlier work (Cockerham, 1969; Weir and Cockerham, 1984) which involved utilisation of an analysis of variance (ANOVA) framework to add levels of population subdivision, and with their effects then extracted as components of variation. AMOVA takes this concept further, by also considering the mutational behaviours of a variety of different markers including sequence data, RFLPs and STRs.

This flexibility is achieved by treating any kind of raw molecular data as a Boolean vector pi, that is, a 1 x n matrix of 1s and 0s, 1 indicating the presence of a marker and 0 its absence. A marker could be a nucleotide base, a base sequence, a restriction fragment, or a mutational event (Excoffier *et al.*, 1992).

Euclidean distances between pairs of vectors are then calculated by subtracting the Boolean vector of one haplotype from another, according to the formula $(p_j - p_k)$. If p_j and p_k are visualized as points in *n*-dimensional space indicated by the intersections of the values in each vector, with *n* being equal to the length of the vector, then the Euclidean distance is simply a scalar that is equal to the shortest distance between those two points. The squared Euclidean distances are then calculated using the equation:

$$\delta_{jk}^2 = (p_j - p_k)' W (p_j - p_k)$$

Equation 3.10 Squared Euclidean Distance

where W is a weighting matrix. By default, it is an identity matrix and does not change the value of the final product. However, W can be a matrix with a number of values depending upon molecular change at different locations in a sequence or phylogenetic tree (Excoffier *et al.*, 1992).

These distances are then calculated for all pairwise arrangements of all Boolean vectors, which are arranged into a matrix and partitioned into submatrices corresponding to subdivisions within the population (Excoffier *et al.*, 1992).

$$D^2 = \begin{pmatrix} \delta_{11}^2 & \delta_{12}^2 & \cdots & \delta_{1k}^2 \\ \delta_{21}^2 & \delta_{22}^2 & \cdots & \delta_{2k}^2 \\ \cdots & \cdots & \cdots & \cdots \\ \delta_{j1}^2 & \cdots & \cdots & \delta_{jk}^2 \end{pmatrix}$$

Equation 3.11. Matrix of Euclidean Distances

They are arranged in such a way that the sub-matrices on the diagonal of the larger matrix are pairs of individuals in the same population, while those on the off-diagonal represent pairs of individuals from different populations. The sums of the diagonals in the matrix and sub-matrices yield sums of squares for the various hierarchical levels of the population.

These sums of squares can then be analyzed in a nested analysis of variance (ANOVA) framework. A nested ANOVA differs from a simple ANOVA in that data are arranged hierarchically and mean squares are computed for groupings at all levels of the hierarchy. This allows for hypothesis testing for between-group and within-group differences at several hierarchical levels (Excoffier *et al.*, 1992; Excoffier, 2001).

A series of variance components are generated from these mean squares where σ_a^2 is the component of variance among individuals within a deme, σ_b^2 is the component of variance among demes within groups, and σ_c^2 is the component of variance among groups within the total population, and where the total variance is given by $\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$.

These variance components can then be used to calculate a series of statistics, analogous to *F*-statistics, called phi-statistics (Φ), which summarize the degree of differentiation between population divisions where Φ_{ST} is a hypothesis about differentiation between the population and its component demes, Φ_{SG}, between demes within a group of demes and Φ_{GT}, between groups within the total population. (Excoffier *et al.*, 1992; Excoffier 2001).

$$\Phi_{SG} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2}$$

$$\Phi_{GT} = \frac{\sigma_a^2}{\sigma^2}$$

$$\Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}$$

Equation 3.12. Phi Statistics

These hypotheses can be tested using the null distribution of the variance components. If the variance of the subpopulations does not significantly differ from the null distribution of the variance of the population, the hypothesis that those subpopulations are differentiated from the larger population would be rejected. This null sample is computed by the bootstrap resampling method.

3.4 Phylogenetic Methods

In addition to the approach based on summary statistics, exploration of the genetic relationships of human populations can also be attempted using a phylogenetic approach. Phylogenetics was primarily developed to construct species trees for taxonomic macro-evolutionary purposes. However, with the increasingly widespread availability of DNA data, phylogenetic techniques have been used to construct so-called gene-trees that display micro-evolutionary relationships between different human populations. There are two fundamental types of phylogenetic analysis, distance matrix methods and discrete methods (which include maximum parsimony and likelihood methods).

3.4.1 Distance matrix methods

Distance matrix methods primarily involve the calculation of genetic distances, and the conversion of a matrix of these distances to some meaningful graphical representation. Genetic distances are defined as the extent of gene difference between populations or species that is measured by some numerical quantity (Nei, 1987; Nei and Kumar, 2000).

Distances such as D_S (Nei and Roychoudhury, 1974) and D_{ps} (Bowcock *et al.*, 1994), represent just two of the many measures that have been used. Nei's standard distance, D_S, is a commonly used genetic distance and is defined as follows:

$$D_S = -ln(J_{XY} / \sqrt{J_X J_Y})$$

Equation 3.13 Nei's Standard Genetic Distance

where J_x and J_y are the average homozygosities across loci in populations X and Y.

Measures such as D_S are based either on the product of frequencies of all alleles at shared loci between populations, or on the proportion of all alleles at all shared loci. They do not directly involve a mutation rate over a long period and hence do not necessarily indicate an evolutionary relationship, but are simply a measure of the effect of genetic drift on population genetic diversity based on WFN assumptions.

With the development of different genetic markers including STRs, SNPs and mitochondrial hypervariable regions (HVRs), different distances have been formulated that accommodate the mutational behaviours that each class of marker exhibits, and thus also the underlying evolutionary relationships. For example, there are many formulae for calculating genetic distances between DNA sequences. These formulae display increasing complexity depending on how the mutational behaviour of the sequences is modelled. The basic sequence distance is the Jukes-Cantor distance, which is basically a corrected percentage of the number of nucleotides by which two sequences differ and where the correction is simply a constant to allow for multiple substitutions at the same site. For nucleotide sequences, this constant is set at a value of ¾ (Nei and Kumar, 2000).

From this distance, research has led to the derivation of increasingly complex distance measures, as more is discovered about mutational behaviour in DNA sequences. One example is the Tamura-Nei distance with gamma correction, a distance specially developed for data from the human mtDNA control region (Nei and Kumar, 2000). It is a complicated distance measure with different transversion and transitional mutation rates for each nucleotide represented by a gamma distribution, and thus allows for uneven mutation rates.

Distances can also be calculated from other genetic markers. For example, to use STR marker data to calculate genetic distance, one must understand the behaviour of the addition or subtraction of whole repeats. It has been found that the

mutational behaviour of these markers is best represented by the Stepwise Mutational Model (SMM), first defined by (Kimura, 1968) from allozyme data. This model allows single step mutation and high levels of homoplasy. The most widely used distance for STR data utilising this model is the delta mu squared distance (Goldstein *et al.*, 1996; Goldstein and Pollock, 1997).

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

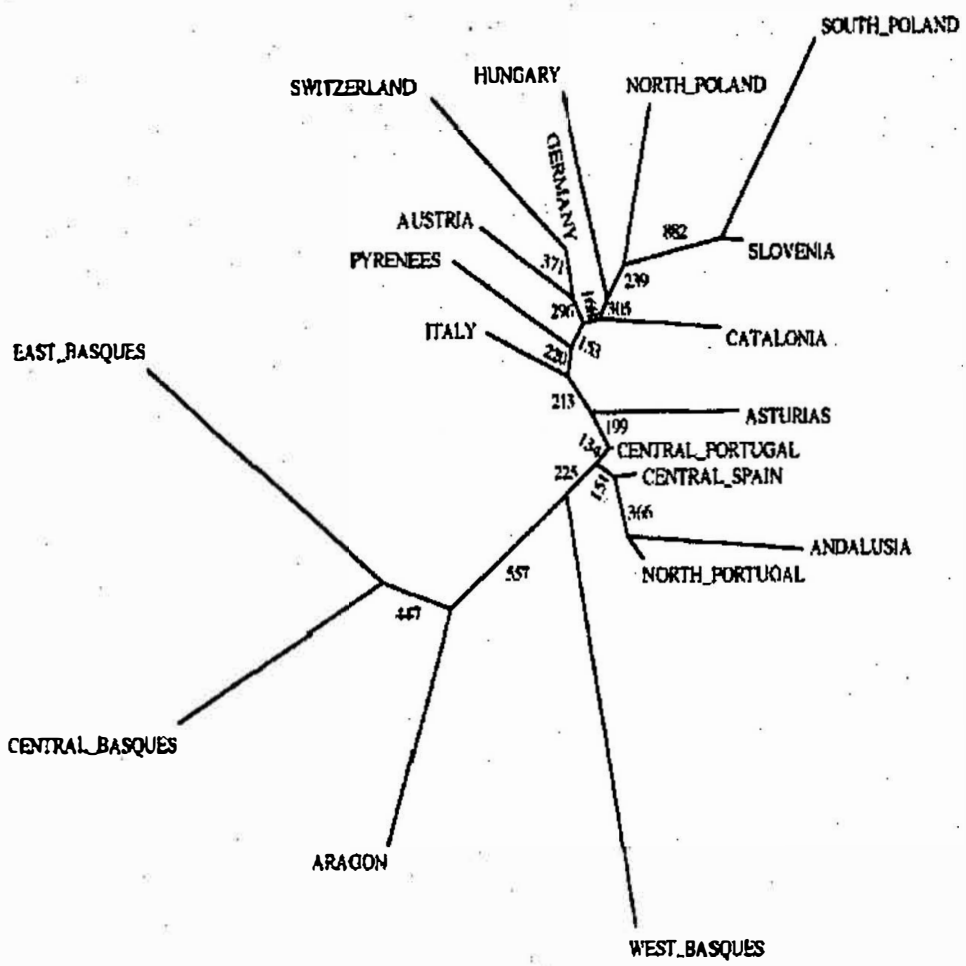
Equation 3.14 Goldstein's Delta Mu Squared Genetic Distance

where μ_A and μ_B are the means of allele size, summed over all loci, in populations A and B respectively. This distance measure is quite accurate in measuring evolutionary genetic distance as it maintains a linear relationship to time over a long period, for example, several thousand generations (Takezaki and Nei, 1996).

3.4.1.1 Drawing the trees: the Neighbour-Joining algorithm

Once a matrix of distances has been calculated, traditionally the next step is to display them in tree form. The method popularly used in molecular anthropology is the Neighbour-joining tree-drawing algorithm. Neighbour–joining (NJ) was introduced by (Saitou and Nei, 1987) as an improvement on a previous method, UPGMA (unweighted pair-group method using an arithmetic average). A pair of neighbours is defined as two populations connected through a single node in an unrooted, bifurcating tree. The Neighbour–joining process begins with a star-like phylogeny. From this initial phylogeny, neighbours are defined as the pair of populations that, when joined, result in the tree of shortest length. The procedure comprises the identification of neighbours among a reduced set of populations until just three combined populations remain, and at that point there is only one unrooted tree. Typically, tree reconstruction is repeated many times from re-sampled data using the bootstrap method (usually 1000 iterations), and a consensus tree is constructed from the most commonly occurring branches. An example of this process is illustrated in Figure 3.1.

Figure 3.1 Neighbour-joining consensus trees for various European populations generated from autosomal STR data



The numbers represent the number of times each node appears in 1,000 bootstrap iterations (Iriondo *et al.*, 2003).

As a distance matrix tree drawing method, NJ is dependant on the evolutionary model used in generating the distance matrix. These models are based on the individual qualities of the marker utilised and, in effect, they are a summary that best fits the observed behaviour of the markers based on data sets. Not all information in a data set is necessarily accounted for in the final distance matrix, and so any tree constructed from it may not be the 'true' tree. However, since they are based on mathematical evolutionary models, trees constructed based on distance matrix methods can be statistically assessed.

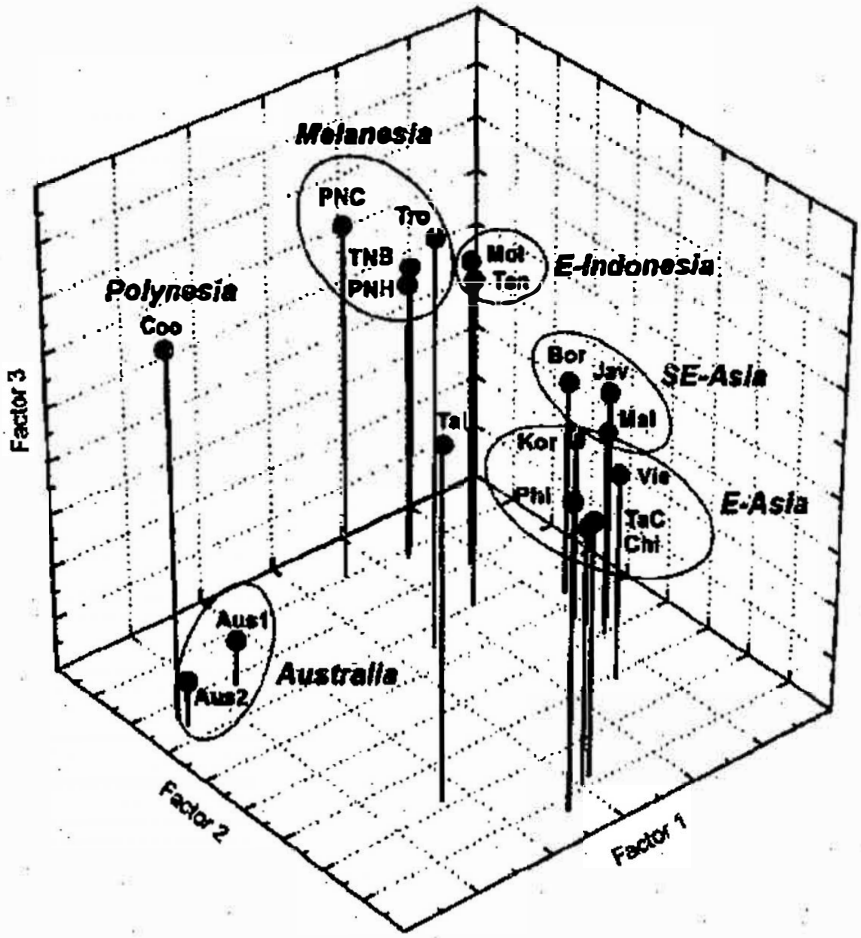
3.4.1.2 Ordination methods

An alternative to representation of the distance matrix as a tree is to use principal components analysis and related clustering methods. First developed by (Hotelling, 1933) and applied to human gene frequencies by (Cavalli-Sforza *et al.*, 1994), Principal Component Analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables termed principal components. The first principal component accounts for the maximum possible degree of variability in the data, with each succeeding component similarly accounting for as much of the remaining variability as possible.

Technically, the principal components of a data set are the eigenvectors obtained from an eigenvector-eigenvalue decomposition of the distance matrix. The eigenvalue corresponding to an eigenvector represents the amount of variability explained by that eigenvector, and the eigenvector of the largest eigenvalue is the first principal component. The eigenvector of the second largest eigenvalue is the second principal component and so on. The result is displayed as a scatter plot where the axes are rotated orthogonally to maximise the variance of the data set (as illustrated in Figure 3.2).

In recent studies, such as those reported by (Zerjal *et al.*, 2002), multidimensional scaling analysis (MDS) has been employed instead of PCA. Broadly speaking, MDS can be considered as a non-parametric alternative to principal components analysis. MDS uses rank order information to identify similarity in a data set and produce a 'map', which is simply a scatter plot that at first glance is similar to a PCA scatter plot. However, as MDS is a non-parametric method, it has the advantage of not incorporating an assumption of normality. MDS methods have only begun to be explored quite recently, as they are much more computationally intensive than PCA.

Figure 3.2 3D PCA plot of Y-chromosome haplotype frequencies from selected populations in Asia, Melanesia and Australia (Kayser *et al.*, 2001a)



3.4.2 Discrete Methods

A phylogenetic alternative to calculating distances is to utilise discrete methods. Discrete methods operate directly on sequences/markers, rather than just pairwise distances, in an attempt to avoid the loss of information that occurs when sequences/markers are converted into distances. The two major discrete methods are maximum parsimony and maximum likelihood (Page and Holmes, 1998).

3.4.2.1 Maximum Parsimony

Maximum parsimony methods take explicit notice of the character values observed (usually nucleotides) for each population, rather than working with distance. They have the feature that evolutionary model assumptions do not need to be specified (Weir, 1996). Maximum parsimony is the method that has been used to define human mtDNA haplogroup and Y-chromosome SNP genealogies (Macaulay *et al.*, 1999; Underhill *et al.*, 2000).

In maximum parsimony, all possible tree topologies are generated based on the data of interest. For each possible topology, the sequences at each node are inferred to be those that require the least number of changes to generate each of the two immediately descendant sequences/haplogroups. The total number of changes required to traverse the whole tree is then determined. The tree with the minimum total of changes is the most parsimonious and thus is the final tree (Weir, 1996). The major disadvantage of the method is that it is not based on statistical principles, so the results cannot be examined with statistical tests (Weir, 1996). Also, in trying to minimise the number of changes it is implicitly assumed that multiple events are improbable. Therefore in cases where a large amount of change has occurred, parsimony methods can generate trees that may prove to be incorrect as more data are added.

3.4.2.2 Maximum Likelihood

The method of maximum likelihood attempts to reconstruct a phylogeny using an explicit model of evolution, where the likelihood of observing a given group of sequences/markers is maximised for each topology and the topology with the highest maximum likelihood is chosen as the true tree. The basic model used in maximum-likelihood phylogeny has two parts: a branching process which describes

how populations diverge over time, and a sampling model which describes the internal properties of the populations. The sampling model is much harder to formalize than the branching process, so it is often ignored; the populations being studied are assumed to be chosen independently at random. If M is a parametric model of the evolutionary branching process, and S is a set of observed populations, maximum likelihood phylogeny methods attempt to find a tree T that maximizes the likelihood:

$$L(T) = P(S \mid T, M)$$

Equation 3.15 Likelihood of a Phylogenetic Tree

The method uses more information in a population set than either distance or parsimony methods, but it is also the slowest and most complex to generate. These complexities are further explored in Section 3.5 in relation to the Coalescent.

3.4.3 Network Analysis

Previous phylogenetic methods were initially developed to determine the pattern of inter-species descent, and not descent within a single species. Thus it could be argued that using phylogenetics to construct a tree of human evolution is inherently flawed. To understand the problem in modelling genealogy within a species, the specific characteristics of different evolutionary relationships must be understood. Relationships between genes sampled from different species are hierarchical, as they are a product of reproductive isolation and population fission over long timescales, during which time mutation combined with population divergence could lead to the fixation of different alleles and ultimately to non-overlapping gene pools.

By contrast, relationships within a species are not hierarchical because they are the product of sexual reproduction, smaller numbers of relatively recent mutations, recombination and overlapping gene pools (Posada and Crandall, 2001). Traditional methods such as NJ and maximum parsimony cannot allow for the fact that several phenomena violate some of their assumptions, such as gene flow between populations. These effects can cause the existence of multifurcations (multiple descendant haplotypes) and reticulations (the result of homoplasy, recombination and hybridization) (Posada and Crandall, 2001).

One method that does allow for these effects is network construction. Networks allow for persistent ancestral nodes and multifurcations, which are common events in intra-species genealogies, and they are designed to show all possible trees rather than one consensus tree. Two networking methods predominate, median networks and median-joining networks.

In median network analysis, whether based on DNA sequences, RFLPs, SNPs or STRs, the information is converted to binary data and constant sites are eliminated. Sites that support the same split are grouped in one character, which is weighted by the number of sites grouped. This leads to haplotypes being represented as 0-1 vectors. Median or consensus vectors are then calculated for triplets of vectors until a median network has been completed (Bandelt *et al.*, 1995; Posada and Crandall, 2001). In such an analysis it is not uncommon that a large number of haplotypes are generated, resulting in networks which are impractical to display. Therefore the reduced median network was developed, where the raw median network is subject to predictions based on frequency information (see section 3.5), and only the most parsimonious trees are guaranteed representation (Bandelt *et al.*, 2000).

The median-joining method is a further refinement of the median network methodology. In median-joining all possible minimum spanning trees (MSTs) are generated and combined into a single network. MSTs are a graph theory construct that connects n haplotypes in a complete network of $n-1$ branches, with the total length of the branches being the minimum necessary to connect all haplotypes (Posada and Crandall, 2001).

Once the MSTs are combined into a single network, median vectors are added to the network via the parsimony criterion. These vectors represent missing intermediates between haplotypes. The method results in a practical network which is fast to generate and, unlike reduced median networks, can involve multistate data (Bandelt *et al.*, 1999; Posada and Crandall, 2001). It is now widely applied in Y-chromosome and multistate mtDNA analysis (Underhill *et al.*, 2000; Torroni *et al.*, 2001).

Many molecular anthropological studies involve the creation of networks that link all of the known haplogroups. Network methods have the advantage of enabling

the researcher to infer divergence times without needing to account for the effect of population size. An example is the statistic ρ that is directly estimated from a network and is the average number of sites differing between a group of sequences (or loci) and a specified common ancestor (Forster *et al.*, 1996). This parameter can be compared to the statistics π, the average pair-wise sequence difference, and *S*, the number of segregating sites (i.e. positions variable in the sample sequences), and is able to infer relationships between populations of interest, e.g., the existence of subdivision and/or past population expansion (Forster *et al.*, 1996).

3.5 Coalescent Theory

With the advent of modern computational technologies, methods have been developed which combine the advantages of both summary statistic and phylogenetic methods, based on the concept called the Coalescent. Considered as the culmination of the WFN model, Coalescent theory was described independently by Kingman, (1982); Hudson, (1983) and Tajima, (1983). However, it is Kingman's explanation of the Coalescent that has become widely adopted.

There are two fundamental tenets of the Coalescent. The first is that since, by definition, selectively neutral variants do not influence reproductive success, it is possible to separate the mutation process and the genealogical process, in effect separating state from descent. Therefore the allelic states of any group of individuals can be generated by assigning an allelic state to their most common recent ancestor (MRCA). The lines of descent can then be followed forward in time, using the rule that offspring inherit the allelic state of their parents unless there is a mutation, which occurs with some specified probability in each generation.

The second tenet is that it is possible to model a genealogy of individuals backward in time without referring to the rest of the population. This can be viewed as randomly 'picking' their parents as one goes back in time. Whenever two lineages 'pick' the same parent, their lineages coalesce. Eventually all lineages coalesce into a single lineage, the MRCA, from which the time to the most recent common ancestor (TMRCA) can be estimated.

Using this approach, the Coalescent can be modelled with efficiency. Formally, the probability that two haploid sequences in the current generation came from a single ancestral sequence is given by $1/N$, where N is the effective population

size. Furthermore, the probability of coalescence occurring t generations ago is the probability of not coalescing for the first t-1 generations, times the probability that a Coalescent event occurred in the *t*th generation. This is given by the geometric distribution:

$$P(\text{Coalescence at t}) = \frac{1}{N}(1-\frac{1}{N})^t$$
$$\approx \frac{1}{N}e^{-(1/N)t}$$

Equation 3.16 Probability of Coalescence at Time t

Extending this logic to *n* possible sequences gives

$$P(\text{Coalescence}) = \frac{n(n-1)}{2N_e}$$

Equation 3.17 Overall Probability of Coalescence

where N_e is the expected population size. From this equation the expected time to coalescence (or TMRCA) is intuitively, $E[T_n] = 2N_e / n(n-1)$. It can be seen that the primary difference between traditional phylogenetic methods and Coalescent theory is that coalescence methods can be conceptualised as the 'reverse' of phylogenetics. This is because phylogenetic methods go forward in time to define a time of divergence in a genealogy, while in Coalescent theory one searches back in time to seek the time of coalescence.

Computationally, the Coalescent can be viewed as a Markov continuous-time process and thus modelled accordingly. Noting that Coalescent time runs backward, let $t_0 = 0$ denote the present, corresponding to the leaves of a genealogical tree and let t_j, $j \in \{1, 2, \ldots, n-1\}$, denote the time of the *j*th most recent coalescence event. Then t_{n-1} denotes the time of the MRCA, corresponding to the root of a genealogical tree. Under the standard coalescent model the between-coalescence intervals $t_j - t_{j-1}$ can be defined as having independent exponential distributions:

$$P(t_j > t \mid t_{j-1} = t') = e^{\beta_j (t'-t)} \qquad \beta_j = \binom{n+1-j}{2} \qquad \text{For } t > t'$$

Equation 3.18 Probability of Coalescence Occurring in a Given Time Interval (Wilson et al., 2003)

Furthermore, mutations in the standard model can be imagined to occur along the branches of a genealogical tree at the points of a homogenous Poisson process with rate $\theta/2$, corresponding to a mutation rate of $\mu = \theta/2N$ per locus per generation in a population of N sequences (Wilson *et al.*, 2003). This model can be further expanded to include factors such as population growth and population splitting (Marjoram and Donnelly, 1994; Donnelly and Tavare, 1995; Beerli and Felsenstein, 2001).

3.5.1 Likelihood analysis and the Coalescent Theory

By defining the Coalescent, it is not only possible to estimate the TMRCA but also the parameters of what is an essentially random genealogical process. This is best illustrated via likelihood analysis. The equation for phylogenetic likelihood inference is given by:

$$L = P(D \mid G, \mu)$$

Equation 3.19 Phylogenetic Likelihood

where L is the likelihood, D are the data, G represents the genealogical process (i.e., the tree) and μ is the chosen mutation model. The objective of the analysis is to estimate the maximum likelihood of parameter G and thus estimate time of divergence. In contrast, the analogous equation in a Coalescent setting is:

$$L = \sum_G P(D \mid G, \alpha) P(G \mid \alpha)$$

Equation 3.20 Discrete Coalescent Likelihood

where α is the collection of parameters (including, but not restricted to θ) for the genealogical process, and the objective of the analysis is typically to estimate these parameters. The tree or genealogy, G, becomes a so-called nuisance parameter, which in theory can be removed by averaging out the likelihood over all possible values.

This is a more powerful statistical model than the traditional phylogenetic likelihood model, as the values and effects of variables such as population size and migration on the coalescence process can be assessed. While providing the potential of greater in-depth analysis, the major drawback of likelihood Coalescent analysis is its mathematical complexity, as summing over all possible genealogies is exceedingly difficult, if not impossible (Rosenberg and Nordborg, 2002). For example, using equation 3.20 it can be calculated that in a sample of just 10 chromosomes, 2,571,912,000 tree topologies (G) can be generated. In addition, since the branch lengths of the trees are generally continuous as they represent time, the following better represents the likelihood:

$$L = \int_G P(D \mid G, \alpha) P(G \mid \alpha)$$

Equation 3.21 Continuous Coalescent Likelihood

This equation makes the exact computation of the likelihood even less possible, since integration must be undertaken over the space of all possible G, which rapidly approaches infinity.

3.5.2 Markov Chain Monte Carlo

While the likelihood cannot be directly calculated, it is possible to generate an analytical approximation via Monte Carlo Markov Chain (MCMC) simulation. This technique involves solving the complex equation numerically using Monte Carlo integration, while keeping the variables generated dependant on the desired model through Markov Chain stochastic iteration until the Markov chains converge to a stationary distribution (Metropolis *et al.*, 1953; Hastings, 1970).

In explaining MCMC, some prior definitions are required. Say X is a vector of k random variables with distribution $\pi(.)$, where the current distribution $\pi(.)$ is the

likelihood (Equation 3.21). The main objective of the analysis is to evaluate the expectation of some function of X, $E[f(X)]$, where:

$$E[f(X)] = \frac{\int f(x)\pi(x)dx}{\int \pi(x)}$$

Equation 3.22 Expectation of a function of X

One approach to solving this equation is to use a numerical integration technique called Monte Carlo integration. In this technique, samples $\{X_t, t=1,\ldots,n\}$ are taken from $\pi(.)$ and used to approximate $E[f(X)]$ where:

$$E[f(X)] = \frac{1}{n}\sum_{t=1}^{n} f(X_t)$$

Equation 3.23 Approximation of the Expectation of a Function of X

When the samples $\{X_t\}$ are independent, the laws of large numbers ensure that this approximation can be made as accurate as desired by increasing the sample size n.

For the coalescent model the problem with this approach is that drawing $\{X_t\}$ independently from $\pi(.)$ is not feasible as $\pi(.)$ is non-standard, i.e., it is not symmetric. So a simple random sampling of variables from $\pi(.)$ would result in a biased sample that is not representative of the true distribution. However, $\{X_t\}$ need not necessarily be independent as long as it is generated by a process which draws samples through the support of $\pi(.)$ in the correct proportions. One way to do this is by generating a Markov chain, (the second "MC" in MCMC) having $\pi(.)$ as its stationary distribution.

A Markov chain can be defined as follows: Suppose we generate the sequence of random variables $\{X_0, X_1, X_2, \ldots\}$ where at each time $t \geq 0$, the next state X_{t+1} is sampled from a distribution $P(X_{t+1}|X_t)$ which depends only on the current state of the chain X_t. This is a Markov chain where $P(.|.)$ is defined as the transition kernel. The important property of such chains is that after the starting state, X_0,

which is subject to regularity conditions, the chain gradually forgets this initial state and will converge to a unique stationary distribution $\phi(.)$. As t increases, the sampled points $\{X_t\}$ will look increasing like dependant samples from $\phi(.)$. It is then possible to use the output of the Markov chain to estimate $E[f(X)]$, where X has the distribution $\phi(.)$, as:

$$\tilde{f} = \frac{1}{n-m}\sum_{t=m+1}^{n} f(X_t)$$

Equation 3.24 Ergodic Average

with f defined as the Ergodic Average, and m defined as the number of burn-in iterations where these samples are discarded and calculation of the average starts from iteration $m+1$.

In MCMC, the aim is to construct a Markov Chain such that its stationary distribution $\phi(.)$ converges to our distribution of interest, $\pi(.)$. The core method used to generate such a chain is the Metropolis-Hastings algorithm. The kernel of this algorithm is the acceptance probability, where at each time t, the next state in the chain, X_{t+1}, is chosen by first randomly sampling candidate point Y from a proposal distribution $q(.|X_t)$. The candidate point Y is then accepted with probability $\alpha(X_t, Y)$:

$$\alpha(X,Y) = \min(1, \frac{\pi(Y)q(X \mid Y)}{\pi(X)q(Y \mid X)})$$

Equation 3.25 Metropolis-Hastings Acceptance Probability

If the candidate point is accepted, the next state becomes $X_{t+1} = Y$, but if it is rejected the chain does not move, i.e., $X_{t+1} = X_t$. Remarkably, it is shown that any proposal distribution, $q(.|.)$, will ultimately deliver samples from the distribution of interest, $\pi(.)$. The rate of convergence to this distribution depends on the relationship between $q(.|.)$ and $\pi(.)$. Choosing a $q(.|.)$ which a) converges, and b) converges in an acceptable time frame, can become more of an art rather than a science.

3.5.3 Maximum Likelihood estimation of parameters using MCMC

If an adequate proposal distribution can be constructed, the use of MCMC in genetic studies is, however, very powerful. One example is the maximum likelihood method for the estimation of migration rates and population sizes devised by (Beerli and Felsenstein, 1999, 2001). In these studies a likelihood model was proposed, equivalent to equation 3.20 where G are the genealogies, D the data and P the parameters of interest:

$$L(D \mid P) = \sum_{G} P(D \mid G) P(G \mid P)$$

Equation 3.26 Likelihood Model of Beerli and Felsenstein (1999, 2001).

The authors wanted to explore two specific parameters, migration rate and effective population size. While the model involving these two parameters was developed, there was still the problem of their estimation from the likelihood sample space (Beerli and Felsenstein, 1999). To calculate the likelihood, they constructed an approximation using the Metropolis-Hastings approach. Instead of evaluating the likelihood, $L(P)$ as an approximation of the maximum value of the likelihood, they evaluated the ratio of likelihoods $L(P)/L(P_0)$ where P_0 are randomly chosen values of the parameters previously used to generate genealogies G_i:

$$\frac{L(P)}{L(P_0)} \approx \frac{1}{g} \sum_{i=1}^{g} \frac{P(G_i \mid P)}{P(G_i \mid P_0)}$$

Equation 3.27 Likelihood Ratio Approximation

In turn, this ratio was approximated using the Metropolis-Hastings method, where the acceptance probability of a change to a newly found genealogy after a migration event, G_n is:

$$\alpha(G_n, G_0) = \min(1, \frac{P(D \mid G_n)}{P(D \mid G_0)})$$

Equation 3.28 Metropolis-Hasting Acceptance Probability for Equation 3.27

where D = the data and G_0 is a previously generated genealogy before the migration event. If the genealogy G_n is accepted, it is the starting point of a new G_0. If not, the previous one is still used. After the generation of a large number of genealogies and manipulation of the conditions to achieve convergence), the proposal distribution $P(D|G_n)/P(D|G_0)$ converges to $L(P)/L(P_0)$, i.e., the approximation of the maximum value of $L(P)$, thus generating an approximation of the parameters of interest.

3.5.4 Bayesian analysis and the Coalescent Theory

An alternative to the likelihood approach is to expand it to a full Bayesian analysis. Bayesian analysis is the branch of statistical theory aimed at incorporating prior information into the estimation process. The theory is named after the Reverend Thomas Bayes (1702–1761). While the theory is almost three hundred years old, its applicability was only realised during the latter half of the twentieth century due to the high computational demands it imposes.

If we define D as the data and α as the set of parameter values to be estimated, Bayes' theorem states:

$$P(\alpha \mid D) = \frac{P(\alpha)P(D \mid \alpha)}{P(D)}$$

Equation 3.29 Bayesian Theorem

That is, the probability of the parameters given the data, P(α | D), is equal to the prior probability of the parameters (obtained from previous experience), P(α), times the likelihood, P(D|α), divided by the total probability of the data, P(D), which is obtained by summing the product of the prior probability and the likelihood over all values of α. The probability P(α |D) is more commonly referred to as the posterior probability. An interesting point, as seems intuitive by inspection of Equation 3.29, is that R.A. Fisher developed likelihood inference from inspecting Bayes Theorem. Fisher used the likelihood component of the theorem in isolation, as

he did not agree with the philosophy of Bayesian inference but nonetheless inferred a use for likelihood and developed maximum likelihood inference (Box, 1978).

The difference between likelihood inference and Bayesian inference is that, in a Bayesian world, parameters are drawn from a higher probability distribution. This method of inference therefore requires a decision on how much is known about parameters before the analysis, which in turn reveals how much knowledge has been gained from looking at the data. In likelihood inference, parameters have fixed unique values of which it is assumed nothing is known, and the goal is to represent belief in the different possible values given in the data.

As with the approximation of the maximum likelihood in likelihood models, the posterior distribution of Bayesian models is approximated via the Metropolis-Hastings algorithm. But due to the incorporation of priors, given P is a vector of parameters, P' is the new candidate vector and D are the data, the acceptance probability, is as follows:

$$\alpha(P,P') = \min(1, \frac{\pi(P')}{\pi(P)} \frac{q(P \mid P')}{q(P' \mid P)} \frac{p(D \mid P')}{p(D \mid P)})$$

Equation 3.30 Bayesian Metropolis-Hasting Acceptance Probability

where q(.|.) is the proposal distribution, p(.|.) is the likelihood, and π(.) is the prior density of the parameters.

The advantage of a Bayesian approach to coalescent analysis is that it allows the factoring of background information via prior distributions into the model. Therefore prior knowledge of factors such as population bottlenecks can be included. For example, if it is assumed that N_e in humans has been approximated as between 5,000 and 50,000, this parameter could be represented by a uniform prior distribution where all values between 5,000 and 50,000 are equally likely. This would result in the posterior representing a narrower range of possible values, and thus having a higher statistical power rather than a likelihood model that covers all possible outcomes and produces a less powerful result.

The software package *BATWING* (Wilson *et al.*, 2003) approximates parameters such as mutation rate and population size and the variable TMRCA using

the Bayesian approach, and so can allow researchers to investigate the effect of parameters such as population growth rates, mutation rates and population sizes using the one model. This flexibility is taken even further by the software package *BEAST* (Drummond *et al.*, 2002), which uses an input format based on XML (eXtensible Markup Language (World Wide Web Consortium, http://www.w3.org/XML) that allows the user to design and run a large range of molecular sequence inference models using the one program. The advantage offered is that the developers of *BEAST* do not have to predict every analysis that researchers may wish to perform, and so be required to explicitly provide an option for all potential analyses. Conversely, the flexibility also means that it is possible to construct models that do not perform well under the MCMC inference framework used.

3.5.5 Bayesian population assignment

The revolution in the development of Bayesian analysis in genetics through the development of the Coalescent has led to new possibilities in analysing population genetic data. One example is the Bayesian model-based population assignment algorithm developed by Pritchard *et al.* (2000) which identifies subgroups that have distinctive allele frequencies. This is undertaken by placing individuals into K clusters, where K is chosen in advance but can be varied across independent runs of the algorithm. Individuals can have membership in multiple clusters, with membership coefficients summing to 1 across clusters.

The basic model developed by (Pritchard *et al.*, 2000) is summarized as follows. Suppose N diploid individuals who are assumed to originate in one of K populations, each with its own set of characteristic set of allele frequencies, are genotyped at L loci. Let X be a vector of the genotypes of individuals, Z a vector of the unknown population of origin of the individuals, and P a vector of the unknown allele frequencies in all populations, then our knowledge about Z and P can be represented by the posterior distribution:

$$P(Z,P \mid X) \propto P(Z)P(P)P(X \mid Z,P)$$

Equation 3.31 Pritchard's Bayesian Clustering Posterior Probability

To construct the priors and the likelihood, first assume that the vectors described contain the following elements:

$(x_l^{(i,1)}, x_l^{(i,2)})$ = genotype of the ith individual at the lth locus where i=1,2,..., N and l=1,2,..., L

$z^{(i)}$ = population from which individual i originated

p_{klj} = frequency of allele j at locus l in population k, where k = 1,2,...,K *and* j = 1,2,...,J

where J_l is the number of distinct alleles observed at locus l. From these elements, the likelihood and prior for Z are given by:

$$P(X \mid Z,P) \approx P(x_l^{(i,a)} = j \mid Z,P) = p_{z^{(i)}lj}$$

$$P(Z) \approx P(z^{(i)} = k) = 1/K$$

Equation 3.32 Likelihood of X (a), and the Prior of Z (b)

The prior $P(P)$ is modelled by a Dirichlet distribution, $D(\lambda_1, \lambda_2,...,\lambda_J)$ distributed on allele frequencies $p = (p_1, p_2,..., p_J)$, with the property that these frequencies sum to one. The final model for this prior is as follows:

$$p_{kl.} \sim D(\lambda_1, \lambda_2,..., \lambda_{Jl})$$

Equation 3.33 Approximation of the Prior of P

The posterior distributions for each K are then approximated via MCMC, and the K with the highest probability is deemed to be the correct number of clusters in the given data set. The model assumes no prior admixture, however this relatively simple model was expanded to include Q, a multidimensional vector of ancestry proportions composed of probability vectors $q^{(i)}$, thus allowing individuals to have recent ancestors in more than one population (Pritchard *et al.*, 2000). Therefore the prior for Z becomes:

$$P(Z) \approx P(z^{(i)} = k) = q_k^{(i)}$$

$$where\ q_k^{(i)} \sim D(\alpha_1, \alpha_2, ..., \alpha_{ki})$$

Equation 3.42 a-b Prior of Z for the Admixture Model

This method has been used in a global molecular anthropology study (Rosenberg *et al.*, 2002), which found that clustering of human populations worldwide was in agreement with the OOA theory of modern human origins. However, since that study, the ability to vary the value of α and λ for each population has been added to the model (Falush *et al.* 2003). Furthermore, Falush *et al.* (2003) redefine the prior for *P* allowing for correlation of allele frequencies by creating a multidimensional vector P_A which contains the allele frequencies in a hypothetical 'ancestral' population *p*, where:

$$p_{kl.} \sim D(p_{Al1}\frac{1-F_k}{F}, p_{Al2}\frac{1-F_k}{F}, ..., p_{AlJ_l}\frac{1-F_k}{F})$$

$$where\ p_{Al.} \sim D(\lambda_1, \lambda_2, ..., \lambda_{Jl})$$

Equation 3.35 Prior of P Assuming Correlated Allele Frequencies

The term F_k represents the effective population size of population *k* during the time since divergence, and is termed as such as it has a close correlation with Wright's F_{ST}.

Even with more flexible priors for Z and P, the structure algorithm still has limitations as the effects of recombination and thus linkage are not considered. To correct this oversight, a linkage model has been developed. This model includes assessment of linkage disequilibrium caused by correlation between markers along the same chromosome, as well as correlations among markers across the genome tested for by previous models.

In this 'linkage' model, a chromosome is composed of 'chunks' that are derived, as unbroken units, from one or another ancestral population and all allele copies on the same chunk are derived from the same population (Falush *et al.* 2003). For haploid data, the new prior for Z is:

$$P(Z) \approx P(z_l^{(i)} = k \mid r, Q) = q_k^{(i)} \text{ and,}$$

$$P(z_{l+1}^{(i)} = k' \mid z_l^{(i)} = k, r, Q) = \begin{cases} exp(-d_l r) + (1 - exp(-d_l r))q_k^{(i)} & if\, k' = k \\ (1 - exp(-d_l r))q_k^{(i)} & otherwise \end{cases}$$

Equation 3.36 Prior of Z in the linkage model

Where d denotes the genetic distance from locus l to locus $l+1$ and r is the chromosome breakpoint rate per unit of genetic distance, where breakpoints occur at random as a Poisson process. For diploid data and independently for each individual i, the zs along each of is two chromosomes form independent Markov chains, satisfying the previous equations (Falush *et al.* 2003).

3.6 Summary

The ability to model and simulate populations and the genetic effects that act upon them has allowed an insight into the genetic mechanisms that govern the present structure of human genetic diversity. Nonetheless, an understanding of human genetic diversity requires not just comprehension of the genetic basis of the variation but also a description of its history. Ideally, a description of this type would include the place and time of origin of a genetic marker variant and an explanation of its presence, frequency, and geographical distribution (Goldstein and Chikhi, 2002). By implication, history also includes the genetic mechanisms that could explain the appearance of the variation (such as mutation rate and pattern), the population factors that may explain allele frequency and distribution (migration, population size, and genetic drift), and interaction with the environment (Goldstein and Chikhi, 2002).

Attempting to combine the multifaceted histories of human populations with the emerging complexity of modern population genetic modelling theory presents an imposing challenge. To meet this challenge it is essential to ensure the applicability of the population genetic models to the investigation of human population structure, and the effect of the resultant findings on the ability to detect genealogical processes both recent and ancient. It is the applicability and usefulness of such modelling that forms the major focus of this study.

CHAPTER 4

SUBJECTS AND METHODS

4.1 Introduction

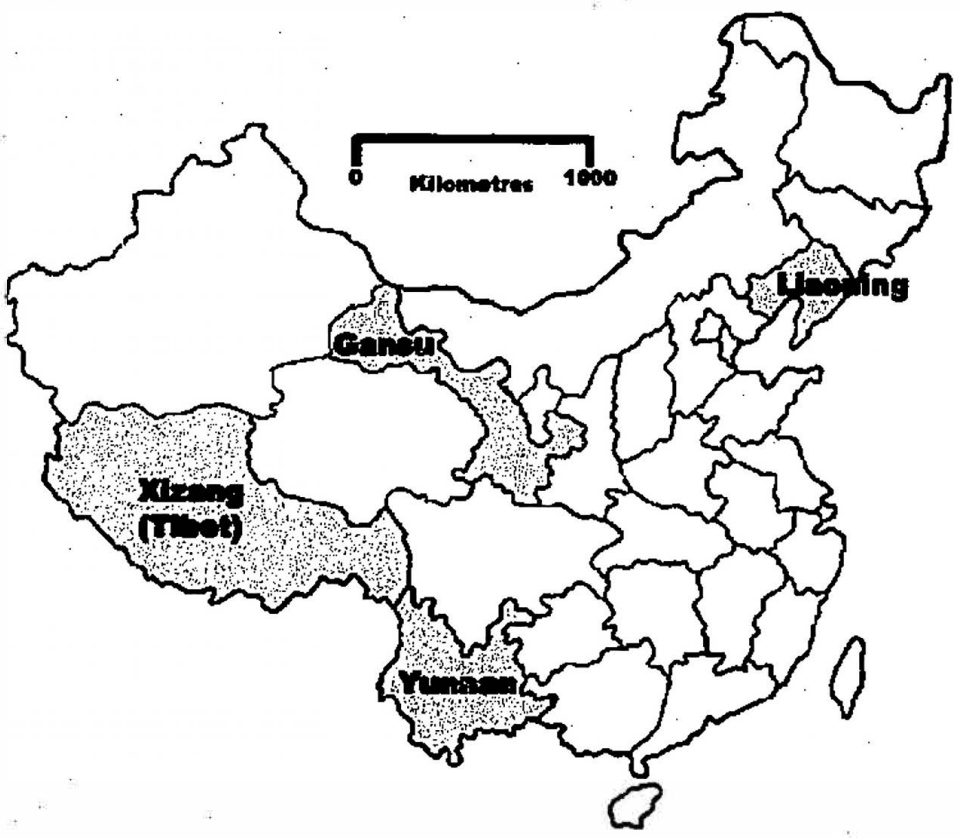
The present study attempts to explain, at least in part, the current patterns of human genetic diversity in China. As alluded to in the Introduction, nine Chinese populations will be investigated; the Han and Hui of Liaoning province, Northeast China; the Bo'an, Dongxiang and Sala of Gansu province, Northwest China; the Tibetan population in Xizang (Tibet) autonomous region, Western China; and the Yao and Miao of Yunnan province, Southwest China (see Figure 4.1). Also included are the Kucong, a population resident in Yunnan province not officially recognised as an ethnic population (*minzu*) by the Government of PR China.

The eight official ethnic populations, or *minzu*, in the study can be separated into two groups on a historical basis comprising populations of 'Ancient' and 'Recent' origins respectively. According to available anthropological and historical records, the ancestry of the Han, Miao, Yao and Tibetan populations can be traced back over several thousands of years in China (Du and Yip, 1993), while the Bo'an, Dongxiang, Hui and Sala are Muslim populations that have existed in China since around 700 AD, soon after the establishment of Islam in the country (Lipman, 1997; Gladney, 1998). This split in population history offers a useful opportunity to assess the effect of migration, population amalgamation and population isolation on genetic diversity in China.

4.2 Historical background of the study populations

The historical chronology of China is most readily discussed in terms of Dynastic periods, since Chinese history prior to the 20th century was mainly classified in this manner. Table 4.1 shows the major Chinese dynasties/periods of the modern era in chronological order. For the subtitles of each of the following sections the official Chinese terminology using the term *minzu* has been adopted, which translated means an ethnic population. In using correct Chinese grammar, the population name and the term *minzu* are combined so, for example, instead of citing the 'Han *minzu*' the correct term is Hanzu.

Figure 4.1 Map of China



The map indicates the four provinces in which the nine ethnic study populations are resident. The four provinces, Gansu, Liaoning, Xizang (Tibet) and Yunnan are shaded with names in bold type.

Table 4.1 Chinese dynasties

Xia	~2000BC - ~1600 BC
Shang	~1600BC - ~1028 BC
Zhou	1027 BC - 221 BC
Spring and Autumn Period	770 BC - 476 BC
Warring States Period	475 BC - 221 BC
Qin Dynasty	221 BC - 207 BC
Western Han Dynasties	206 BC - 8 AD
Chi'in Dynasty	8 AD - 25 AD
Eastern Han Dynasty	25 AD - 220 AD
Six Dynasties Period	220 AD - 581 AD
Sui Dynasty	581 AD - 618 AD
Tang Dynasty	618 AD - 907 AD
Five Dynasties Period	907 AD - 960 AD
Song Dynasty	960 AD - 1279 AD
Yuan Dynasty	1279 AD - 1368 AD
Ming Dynasty	1368 AD - 1662 AD
Qing Dynasty	1662 AD - 1908 AD
Republican period	1911 AD - 1949 AD
Peoples Republic period	1949 AD - present

A brief outline of the major Chinese dynasties/periods. The Xia, Shang and Zhou Dynasties are commonly refereed to in many Chinese historical texts. Combined, these dynasties are also referred to as the Huaxia period. During the Zhou dynasty, other historical periods have also been defined. Historically, the Qin Dynasty formed the first unified Chinese nation. Adapted from Paludan, (1998).

4.2.1 The 'Ancient' populations

4.2.1.1 Hanzu

The Han trace their history back to the Huaxia period, which extended from the 21st to the 8th centuries BC. It is believed that the Han were an ethnic group based on the ancient Huaxia of the middle and lower reaches of the Yellow River, who subsequently assimilated with other local and regional ethnic groups (Du and Yip, 1993). The term Han is used collectively to define the majority population in PR China. It was brought into common usage after the fall of the Eastern Han Dynasty, often thought of as a golden age in Chinese history (Du and Yip, 1993). According to the 2000 Census, the Han made up 92% of the population of PR China and they numbered approximately 1190 million (National Bureau of Statistics of China 2002). The Han are resident throughout PR China, but they are most numerous in the more densely populated east of the country. Although the written Mandarin language is uniform, the spoken language differs from Province to Province, with eight recognised major dialects Mandarin, Wu, Gan, Xiang, North Min, South Min, Hakka and Yue (Moser, 1985). Hence, it can be argued that use of the term Han to encompass all of these peoples is more a political convenience than a true measure of ethnicity.

4.2.1.2 Kucong

To the candidate's best knowledge, no historical material is available in English on the Kucong. The name 'Kucong' means 'forest people' in Mandarin. It is thought they were originally a group of itinerant hunter-gatherers in the tropical forests of the Ailaoshan valley in the Yunnan-Burma border region, but in recent times they have largely been resettled in towns and cities in Yunnan Province (Yang *et al.*, 1994). Linguists have identified the Kucong language as within the Lahu (Tibetan-Burmese) language group (Yang *et al.*, 1994). Because of this linguistic affiliation, the Kucong were placed under the Lahu *minzu*. However, in all other cultural aspects the Kucong have little in common with the Lahu (Xiaotong 1991).

4.2.1.3 Miaozu

With a population of more than eight million, the Miao people form one of the largest ethnic minorities in Southwest China (National Bureau of Statistics 2002).

They are mainly distributed across Guizhou, Yunnan, Hunan and Sichuan Provinces and the Guangxi Zhuang Autonomous Region, with small numbers living on Hainan Island and in Southwest Hubei Province (Du and Yip, 1993). Most of the Miao live in tightly knit communities, with a few resident in areas also inhabited by several other ethnic groups.

The term Miao was first used by the Chinese in pre-Qin times, i.e. before 221 BC, for designating non-Chinese groups in the south of the country. However, during the Qin and Han dynasties, the ancestors of the present Miao people lived in the Yangtze River Valley. Since that time the Miao have migrated both to the north and south of China. In the third century AD, the ancestors of the Miao moved west to present-day Northwest Guizhou and south Sichuan Province along the Wujiang River a southern tributary of the Yangtze. In the fifth century AD, some Miao groups moved to East Sichuan and West Guizhou (Du and Yip, 1993).

By the seventh century AD the Miao had ceased to exist as a separate major non-Chinese group until the formation of PR China in 1949, except in the area which is now known as the Province of Yunnan. However, it is recorded that in the 16th century some people termed 'Miao' settled on Hainan Island. Because of these large-scale migrations over many centuries, the Miao have become widely dispersed (Yin, 1994). The samples to be investigated in this study were obtained from Miao communities resident in Yunnan Province.

4.2.1.4 Yaozu

The Yao, with an estimated population of 2.64 million, live in mountain communities scattered over 130 counties in five southern Chinese Provinces and one autonomous region (National Bureau of Statistics 2002). About 70% live in the Guangxi Zhuang Autonomous Region and the remainder are resident in Hunan, Yunnan, Guangdong, Guizhou and Jiangxi Provinces.

Through the course of history, the Yao have had 28 different names (Du and Yip, 1993). The name 'Yao' was officially adopted after the founding of PR China in 1949. Half of the Yao speak the Yao language, which belongs to the Sino-Tibetan language family as the Hmong-Yao linguistic group along with the Miao language, while others use the Dong language. As a result of close contact with Han

populations and the Zhuang *minzu*, many Yao also have learned to speak Mandarin or the Zhuang language.

Historical evidence first places the Yao in northern Hunan Province around the 3rd to 2nd century BC (Du and Yip, 1993). From that time the Yao have had a history of continuous migration, inhabiting all of Hunan by the 10th century AD. They then were forced to shift southwards *en masse* during the 13th–17th centuries AD, settling in Guangdong and Guangxi Provinces, and during the 17th century some Yao migrated further south to Yunnan Province.

The formation of PR China led to the establishment of the Yao as an official *minzu*, and the creation both of the Guangxi Zhuang Autonomous Region and many autonomous prefectures. The Yao population investigated in the present study is from Yunnan Province.

4.2.1.5 Zangzu (Tibetan)

The Tibetans with a population of 5.4 million, mostly live in the Tibet Autonomous Region, Xizang (National Bureau of Statistics 2002). However, there are also Tibetan communities in Qinghai, Gansu, Sichuan and Yunnan Provinces. The Tibetans have a long history and archaeological evidence has indicated a history back to the Neolithic (~10,000 YBP). The Kingdom of Tibet was founded in the 6th century AD and began contact with the Tang Dynasty, culminating in 641 AD with the marriage of the Tibetan King Songzan Gambo to Princess Wen Cheng of the Tang. In 710 AD, King Chide Zuzain married another Tang princess, Jin Cheng. From that time foward, emissaries travelled frequently between the Tang Dynasty and Tibet (Yin, 1994).

The Yuan Dynasty brought Tibet under the rule of the central dynastic government for the first time during the 13th century AD. It established an institution called the Xuanzhengyuan (or Political Council) and placed it in charge of Buddhist affairs and the military and government of Tibet (Yin, 1994).

The Yuan court set up three government offices to govern the Tibetan areas in Northwest and Southwest China, and in Tibet itself. This marked the beginning of the overall control of Tibet by central Chinese authorities, with the appointment of officials and the establishment of an administrative system. The Qing court

appointed a resident High Commissioner to help with local administration in 1728 AD, and set up the Kasha as the local government in 1751 AD (Yin, 1994). When the Qing Dynasty collapsed in 1908 AD, the Tibetans again opted for self-rule until annexation by PR China in 1951.

Due to the nature of the current political situation, historical, demographic and anthropological sources published during the latter half of the 20th century may have to be treated with caution, as many are Chinese in origin and may not reflect the full historical background of Tibet.

4.2.2 The 'Recent' populations

4.2.2.1 Bo'anzu

The Bo'an, sometimes referred to as Bao'an or Bonan, are an Islamic community and one of China's smallest ethnic minorities with a total population of 16,500 people (National Bureau of Statistics 2002). Their language belongs to the Mongolian branch of the Altaic language family and is close to that of the Tu and Dongxiang ethnic minorities (Du and Yip, 1993). The Bo'an use the Han script, as they do not have a written language of their own (Yin, 1994). The Bo'an sampled for this study, as well as the Dongxiang and Sala, live in Gansu Province in Central-West China.

The term 'Bo'an' literally means a security fortress and it is believed that the Bo'an people were so named because of their origins as a security battalion, or Bo'an Ying (Du and Yip, 1993). The Bo'an community appears to have arisen after many years of intermarriage during the Yuan and Ming periods between Mongolians, who settled as garrison troops in Tongren County, Qinghai Province, and the neighbouring Hui, Han, Tibetan and Tu people (Du and Yip, 1993). The Bo'an formerly lived in three major villages in the Bao'an region, situated along the banks of the Longwu River and within the boundaries of Tongren County, Qinghai Province. However, during the reign of the Qing Emperor Xianfeng (1851 AD – 1862 AD) the Bo'an moved to what is now known as the Dahejia region of Gansu Province. The name Bo'an was only specifically recognized when the community was identified as an official *minzu* after 1949.

4.2.2.2 Dongxiangzu

The Dongxiang live in the Ningxia Hui Autonomous Prefecture situated south of the Yellow River and southwest of Lanzhou, the capital city of the North-West Province of Gansu. Half of the population live in Dongxiang Autonomous County, and the remainder are dispersed mainly in Hezheng and Ningxia counties, the city of Lanzhou and the Xinjiang Uygur Autonomous Region (Yin, 1994). The Dongxiang Muslims number some 513,000 individuals (National Bureau of Statistics 2002).

The Dongxiang ethnic minority received its name from the place in which they live - Dongxiang. However, they were not recognized as a minority before the founding of the Peoples Republic in 1949 and previously were known as 'Dongxiang Hui' or 'Mongolian Hui'. The Dongxiang language is similar to Mongolian. Both languages belong to the Mongolian branch of the Altaic language family, but Dongxiang contains quite a number of words borrowed from Mandarin. A majority of the Dongxiang people also speak Mandarin, which now is accepted as their common written language (Du and Yip, 1993).

Historians are divided in their views on the origin of the Dongxiang ethnic minority. Some hold that they are descendants of Mongolian troops posted in the Hezhou area by Genghis Khan (1162-1227 AD) during his journey westward, while other historians believe they are a mixture of many peoples including Mongolian, Han and Tibetan groups (Du and Yip, 1993).

4.2.2.3 Huizu

At 9.8 million, the Hui are the most numerous Muslim minority in China and they account for some 50% of the total Muslim population (National Bureau of Statistics 2002). The Hui are widely dispersed and are resident in 19 of the 21 Provinces of PR China. It is notable that the Hui are recognised as an official minority, despite their lack of a common territory or a common language. Their extensive geographical spread suggests that the Hui may actually comprise many sub-populations, each which could potentially claim minority status in its own right.

The origins of the Hui are diverse, and it is believed that they include individuals whose ancestors originated in pre-Islamic times, from Central Asia, Iran, and the Middle East (Du and Yip 1993). For example, in 61 AD some 10,000 Turkic families from the Central Asian city-states of Samarkand and Bokhara were recorded as settling in the ancient capital of China, Chang'an, now called Xi'an (Du and Yip, 1993). Other recorded migrations to China include the influx of Persian refugees from what is now modern Iran after the fall of the Sassanid Empire in 652 AD. It is widely believed that the Hui community arose from male Arab, Persian and Central Asian traders who settled in China and married local Han women. An Imperial edict during the fifth year of the reign of the Emperor Hongwu of the Ming Dynasty (1372 AD) supported this practice, stating that 'Mongolians and *se mu people* (Hui) are allowed to marry Chinese but not their own kind' (Gladney 1998). In addition, it is believed that a minority of Turkic Uygurs and Mongolians may have also become assimilated into the Hui.

The term 'Hui' was not used to represent a population until the creation of the Huizu by PR China after 1949. In summary, the present Huizu seem to be a population of convenience, that is, a grouping of Islamic individuals who could not be classified into other Muslim *minzu* (Gladney, 1998).

4.2.2.4 Salazu

The Sala, a Muslim population, number 104,500 individuals and live in Qinghai Province and in the Bo'an-Dongxiang-Sala autonomous county of Jisishan, Gansu Province (Du and Yip, 1993; National Bureau of Statistics 2002). Unlike the Bo'an and Dongxiang, who speak a form of the Mongolian language, the Sala language is Turkic in origin and is similar to Uzbek, Turkman and Uygur (Du and Yip, 1993).

It is believed that the Sala originated from an Oghuz tribe of the Seljuk Turks which travelled east from Samarkand in Central Asia during 1370–1424 AD and settled in XunHua county, Qinghai Province. Later some Sala moved to Gansu Province. Over the course of their history, the ancestors of the Sala are believed to have merged with Tibetans, Han Chinese and Mongolians to form the present-day Salazu (Du and Yip, 1993).

4.3 Sample collection

As indicated in section 4.1, the samples obtained were from populations resident in Northeast, South and West PR China. The Han, Hui, Bo'an, Dongxiang, Sala and Tibetan samples were provided as blood spots on Whatman 3MM filter paper and stored at -80^0C in the Centre for Human Genetics at ECU. The Yao, Miao and Kucong samples were provided as extracted DNA from the Institute of Genetics, Chinese Academy of Sciences, Beijing to the Centre of Human Genetics. Ethical approval in China for research utilising the collected samples was granted by Huazhong University of Science and Technology, Wuhan. The requisite ethical approval to undertake research involving human subjects was provided by the ECU Human Research Ethics Committee.

The numbers of DNA samples stored in the Centre for Human Genetics are as listed in Table 4.2.

Table 4.2 Blood and DNA samples held in the Centre for Human Genetics, ECU.

Population	No. of Samples	Male	Female
Han	230	180	50
Kucong	40	40	0
Miao	50	50	0
Yao	24	24	0
Tibetan	54	54	0
Bo'an	80	80	0
Dongxiang	75	47	18
Hui	53	27	26
Sala	80	59	21

4.4 Extraction of DNA from blood spots

The following protocol was adopted in extracting DNA from blood spots from the relevant population samples. The DNA of each individual was isolated from two blood spots, using proteinase K treatment followed by phenol/chloroform extraction and isopropanol precipitation. The blood spots were cut from the filter paper, quartered, and placed in a 1.5 ml microtube with 250 µl 0.1% Triton X-100 and 15 µl 10mg/ml proteinase K at room temperature. The sample was mixed gently for 1 minute before incubation on a heating block at 50^0C for 30 minutes. Following incubation, the sample was gently mixed for 5 minutes before being incubated at 50^0C for an additional 30 minutes. On completion of the second incubation period, 25 µl 10x SET buffer was added. The sample was mixed thoroughly and incubated at 50°C for 30 minutes. 500 µl of 1:1 chloroform/phenol was then added and mixed by inversion for 1 minute. The sample was centrifuged for 30 minutes at 13,000 rpm, and the supernatant transferred to a fresh microtube with waste paper materials excluded. 25 µl of 3M Na acetate pH 4.9 and 250 µl of 100% isopropanol were added to the supernatant, mixed and left overnight at –20°C to precipitate the DNA from the solution.

On the following day, the sample tube was centrifuged for 30 minutes at 13,000 rpm. The supernatant was discarded and the DNA pellet washed with ice-cold 70% ethanol. The sample tube was then centrifuged for 10 minutes at 13,000 rpm. The ethanol was carefully discarded and the pellet was left to dry at room temperature for approximately 1 hour. Finally, the pellet was resuspended in 50 µl autoclaved distilled water.

4.5 Quantitation of DNA concentration

The DNA concentration and sample purity of the samples was analysed with a Beckman DU 640 UV spectrophotometer, by measuring the optical density of a 1:20 dilution of the original DNA solution at wavelengths 260 nm and 280 nm. The spectrophotometer was blanked against the dH_2O used in the dilution of the DNA solution.

Optical density measurement at 260 nm detects nucleic acids while 280 nm detects proteins. The OD_{260}/OD_{280} ratio thus estimates the purity of the nucleic acid

sample. A ratio of 1.6:1 or above indicates satisfactory purity of nucleic acids for amplification. Quantitation of the sample DNA (ng/µl) was obtained using the formula: Total DNA (ng/µl) = OD_{260} x 50 x dilution

A concentration of at least 5-10 ng/µl of DNA in each sample was required for reliable PCR amplification.

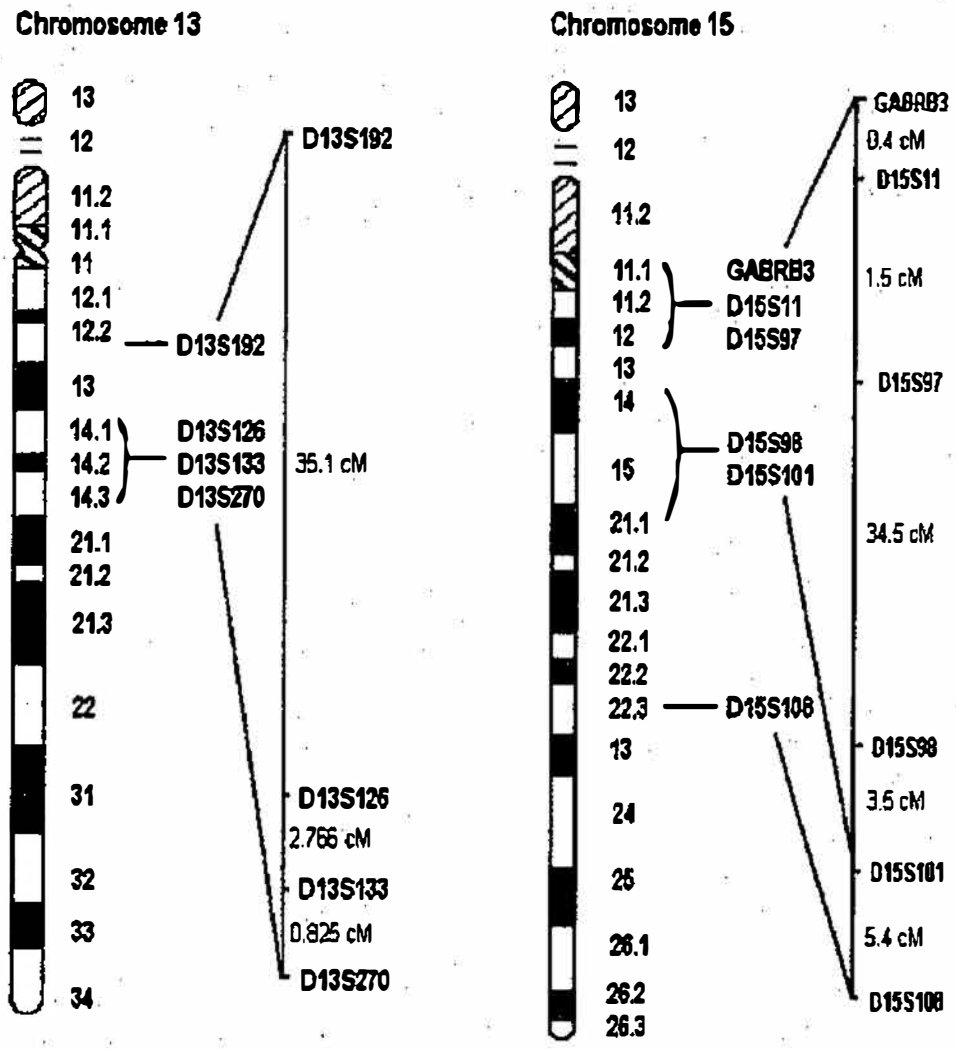
4.6 Polymerase Chain Reaction

Amplification of all sequences in the study was accomplished using the polymerase chain reaction or PCR, an *in vitro* method for synthesising defined sequences of DNA catalysed by a thermostable DNA polymerase enzyme. The reaction consists of three steps: denaturation, annealing and extension. In the first step the DNA is separated into single strands that can be used as a template. Step 2 employs two oligonucleotide primers that anneal to the template DNA at positions flanking the target DNA sequence. Finally, a complementary copy of the region specified by the two primers is synthesised using the enzyme *Taq* polymerase. Repetition of these steps results in exponential amplification of the target DNA sequence.

4.7 Autosomal STR analysis

The autosomal STR markers analysed in this study were chosen from a panel of markers originally selected and recommended by the Department of Genetics, Stanford University. Linkage and cytogenetic maps of chromosomes 13 and 15 indicating the positions of the markers used in the study are shown in Figure 4.2.

Figure 4.2 Comparative linkage maps locating STR markers on chromosomes 13 and 15, with an average distance of ten centimorgans



(adapted from the Stanford Human Genome Centre, 1998)

4.7.1 Autosomal STR PCR conditions

A 10 µl reaction was prepared for each sample of genomic DNA. This consisted of 3 µl of target DNA (5-10 ng/µl), 1 µl of 100 ng/µl forward primer and 1 µl of 100 ng/µl reverse primer (Appendix A1), 1 µl of 10 x reaction buffer (containing 1.5 mM $MgCl_2$ solution [Perkin-Elmer] Appendix B), 0.2 µl of 25 mM $MgCl_2$, 0.5 µl of 5 mM dNTPs, 0.1 µl of *Taq* polymerase, and 3.2 µl of dH_2O. Due to variability in the quality of the target DNA and the fidelity of the primers, the $MgCl_2$ concentration used varied between 1 mM and 4 mM. From these reactions, the requisite markers were successfully amplified in a Perkin-Elmer 9600™ thermocycler following the touchdown PCR protocol detailed in Table 4.3.

Table 4.3 PCR protocol for autosomal microsatellite markers

No. of Cycles	Temperature
1	94°C for 5 minutes (initial denaturation)
15	94°C for 20 seconds (denaturation) 63-56°C for 1 minute (0.5°C decrements) (annealing) 72°C for 30 seconds (extension)
20	94°C for 20 seconds 56°C for 30 seconds 72°C for 30 seconds
1	72°C for 7 minutes (final extension)

Markers D15S98 and D15S97 were found to require lower annealing temperatures, and so a variant of the protocol given above was used, with temperatures of 58-50°C for the touchdown phase and 50°C for the extension phase.

4.8 Agarose gel electrophoresis of microsatellite sequences

To check that the required sequences were amplified, PCR products were tested by agarose gel electrophoresis. This technique employs an electric current to move the negatively charged DNA towards a positively charged electrode through an agarose gel. The larger the allele fragment, the slower it moves through the agarose. Therefore the alleles are differentiated by length fractionation. The PCR products are visualised using ethidium bromide (EtBr) under fluorescent light.

A 3% agarose gel solution (3g agarose powder, Sigma Chemical Company) in 100ml 1xTAE buffer (0.04M Tris-acetate; 0.001M EDTA) was prepared and poured on to an 8.5 cm x 11 cm mini-gel tray, with a small toothed comb inserted at one end of the tray. The gel was allowed to set for approximately 40 minutes at room temperature, after which the comb was removed and the gel was placed into the electrophoresis unit. 3 µl of the PCR products were loaded into the wells with 3 µl of 6x Ficoll loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF; 15% Ficoll (Type 400; Pharmacia). pUC19 DNA/Hpa II (0.5mg/ml; Biotech; fragment size range from 26-501bp) was loaded into lane 1 as a size standard. The gel was electrophoresed at 80V for approximately 20 minutes. It was then stained for 10 minutes in EtBr (1.5µl 100% EtBr, 30ml water) and visualised under a Hoefer® Mighty Bright™ UV transilluminator.

If bands were present, the gel was photographed using the Kodak® DC120 Electrophoresis Documentation and Analysis System™, which included the Kodak® DC Zoom™ Digital Camera and 1D Image Analysis Software™. The resulting digital image could then be stored on a diskette or printed.

4.9 Resolution of STR sequence size through fluorescent labelling

If PCR products were detected using agarose gel electrophoresis, they were then electrophoresed on 6% polyacrylamide gels in the ABI Prism 373A (Applied Biosystems) and sized using an internal standard and the GENOTYPER ™ software (Applied Biosystems). The ABI 373A accurately identifies sequence size via the detection of fluorescent labels on the oligonucleotide primers.

Specifically, the forward primers for the microsatellites are 22-mer oligonucleotides to which either TET (4,7,2,7-tetrachloro-6-carboxyfluorescein), HEX (4,7,2,4,5,7-hexachloro-6-carboxyfluorescein) or FAM (6-carboxyfluorescein) molecules are chemically bonded. These fluorescently labelled primers are then incorporated into the microsatellite markers during amplification with PCR.

For loading into a polyacrylamide gel, 1.5 µl of PCR sample was mixed with 2.5 µl of formamide, 0.5 µl of loading buffer and 0.5 µl of TAMRA-labelled internal size standard. The loading buffer and TAMRA-labelled size standard are supplied in the GENESCAN-500 kit™ (Applied Biosystems). The size standard is a digest of plasmid DNA by the restriction enzymes *PstI* and *BstUI*. The resulting DNA fragments are then labelled with TAMRA (N,N,N,N'-tetramethyl-6-carboxyrhodiamine.

4.10 Detection and quantification of STR markers

A gel mixture of 100 ml 40% acrylamide/bisacrylamide at a ratio of 19:1, 420.5g of urea, and 100 ml of 10xTBE buffer (890 mM Tris-borate, 2mM NaEDTA.$2H_20$), was made up to 1 litre with distilled water and stored at 4°C. 30ml of this gel preparation was mixed with 150 µl of 10% ammonium persulphate and 17 µl of TEMED, dispersed between two glass plates using a 100 ml syringe with a 24 cm well-to-read distance, and fixed together with bulldog clips. A 66 well square tooth comb was placed in the top of the gel apparatus, and the gel allowed to set for approximately two hours at room temperature. The gel was then pre-run for five minutes at 28°C to optimise the temperature of the gel.

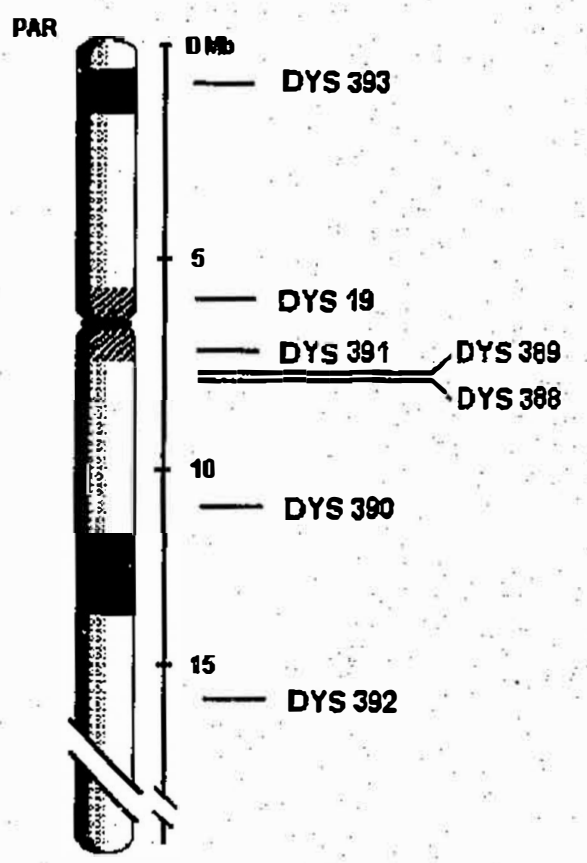
After the pre-run, 3 μl of each sample solution were pipetted into the gel wells and the gel was then run for 8-12 hours using filter B. The use of filter B resulted in the fluorescent markers displaying the following colours: blue (FAM), green (TET), yellow (Hex), and red (TAMRA). The resulting gel pictures were assembled and analysed with the GENSCAN™ software program. The GENOTYPER™ software program was used to analyse the data extracted by GENESCAN™, and to assign peaks to microsatellite alleles. The results were presented as a series of peaks, with each microsatellite marker resulting in either one peak for homozygotes or two peaks for heterozygotes.

4.11 Y-chromosome analysis

The Y-chromosome markers used in this study were chosen from a panel of markers recommended by the Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University (http://www.leidenuniv.nl/fldo). The approximate locations of the Y-STRs used are illustrated in Figure 4.3.

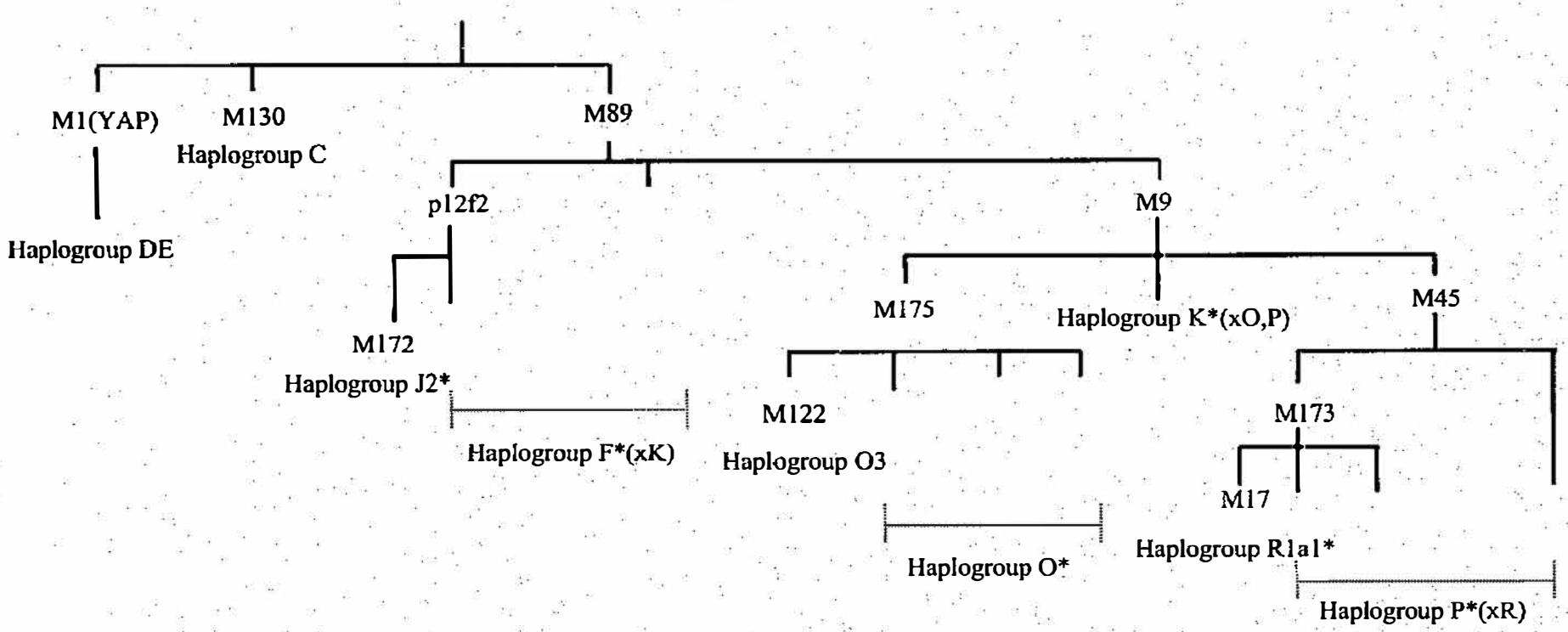
For the current study, samples were analysed for eight unique event polymorphisms (UEPs), consisting of seven Single Nucleotide Polymorphisms (SNPs) and the Alu Polymorphism (YAP) on the Y-chromosome: M1 (YAP), M216/M130 (RPS4Y), M89/M213, M9, M175/M214, M122, M45, and M17 (Underhill *et al.*, 2001). All were detected via RFLP analysis. Unfortunately, due to the limited amounts of sample available, further resolution of markers with MALDI-TOF mass spectrometry was not possible for all populations, and so comparisons were primarily based on these eight key markers. Marker M1 was detected by PCR amplification of either a 455-bp (YAP^+) or 150-bp (YAP^-) fragment that can be resolved by electrophoresis on 2% agarose. A partial genealogy showing the haplogroup designations defined by these UEPs is shown in Figure 4.4.

Figure 4.3 Cytogenetic map of the Y-chromosome indicating the approximate location of the STR markers examined in this study



(Adapted form Bosch *et al.*, 2002)

Figure 4.4 Partial genealogy showing haplogroup designations and the defining mutations of Y-chromosome SNPs



Haplogroups shown are those inferred in the current study and are identified according to the nomenclature of the Y Chromosome Consortium (2002). For comparison purposes Haplogroup DE is defined by the presence of the YAP insertion and also includes markers M96, M174 and M15. Haplogroup C is defined by a derived state for M130 and also includes M217 and M48. F*(xK), defined by a derived state for M89 but ancestral for all other markers, also includes M52, P15 (DYS221) and p12f2. Haplogroup J2 was resolved for samples in Wang *et al.* (2003) and utilised in the current study. It is defined by a derived state in marker M172. Haplogroup K(*xO,P) is defined by a derived state for marker M9 and includes M20 and M46 (Tat). Haplogroup O* is defined by a derived state for M175 but ancestral for M122 and also includes P31 ($ARSEP_{71227}$) and M176 (SRY_{465}). Haplogroup O3, defined by a derived state for M122 also includes M134 and LINE-1. Haplogroup P* is defined by a derived state for M45 and includes M124 and M207 (UTY2), and M173 Haplogroup R1a1* is defined by a derived state for M17 and also includes markers SRY_{10831} and M87.

4.11.1 Y-chromosome STR PCR conditions

A 10 µl reaction was prepared for each sample of DNA. This consisted of 5 µl of target DNA (5-10 ng/µl), 1 µl of 10 x reaction buffer solution (containing 1.5 mM of $MgCl_2$), 1 µl of 25 mM $MgCl_2$ solution, 0.5 µl of 5 mM dNTPs, 0.5 µl of 100 ng/µl forward and reverse primer (see Appendix A2), 0.1 µl of Hot Star™ *Taq* polymerase (Qiagen), and 1.9 µl of dH_20. Two alternative thermal cycling protocols were used depending on the marker being amplified (Tables 4.4 and 4.5). The resultant PCR samples were tested for the presence of product using the method described in section 4.8. Quantitation of marker size was achieved through use of the ABI 373 DNA sequencer™, as described in sections 4.9 and 4.10.

Table 4.4 PCR protocol for markers DYS19, DYS388, DYS392 and DYS393

No. of cycles	Temperature
1	95°C for 15 minutes
5	94°C for 30 seconds 58°C for 30 minutes 72°C for 30 seconds
30	94°C for 30 seconds 54°C for 30 seconds 72°C for 30 seconds
1	72°C for 10 minutes

Table 4.5 PCR protocol for markers DYS389, DYS390 and DYS391

No. of cycles	Temperature
1	95°C for 15 minutes
35	94°C for 1 minute 56°C for 1 minute 72°C for 1 minute
1	72°C for 10 minutes

4.11.2 Y-chromosome biallelic polymorphism PCR

For all UEP markers, PCR reactions were required to amplify the region of interest containing each polymorphic nucleotide. A 25 µl reaction was prepared for each DNA sample. This contained 2.5 µl of 10x reaction buffer solution (containing 1.5 mM of $MgCl_2$), 5 µl of DNA (5-10 ng/µl), 1.5 µl of each primer (10 µM) (Appendix A3), 1 µl of 25 mM $MgCl_2$, 1 µl of 5 mM dNTPs, 0.2 µl HotStar™ *Taq* DNA Polymerase, and 12.3 µl of dH_2O. The program detailed in Table 4.6 was used to amplify each of the SNP markers:

Table 4.6 Y-chromosome biallelic PCR protocol

No. of cycles	Temperature
1	95°C for 15 minutes
15	94°C for 30 seconds 63-56°C for 1 minute (0.5°C decrements) 72°C for 1 minute
30	94°C for 30 seconds 56°C for 1 minute 72°C for 1 minute
1	72°C for 10 minutes

4.11.3 YAP and RFLP assays

M1 (*YAP*): The amplified products were electrophoresed on an agarose gel to differentiate the ancestral from the derived allele. The M1 ancestral allele does not contain the Alu insertion, thus the amplified product is 150bp in size, i.e., it is 291bp smaller than the derived allele (441bp).

M130 (C→T): A 20 µl reaction was set up for each sample, containing 3 µl of amplified product, 2.5 units of *Bsl* I (New England Biolabs), 2 µl of 10 x NE buffer 3 and made up to total volume with dH_2O. The reaction mix was incubated at 37°C for four hours. Digestion of the sample indicates the derived allele (T).

M89/213 (T→C): A 20 µl reaction was set up for each sample, containing 3 µl of amplified product, 5 units of *Nla* III (New England Biolabs), 2 µl of 10x BSA

(New England Biolabs), 2 µl of 10x NE buffer 4 and made up to total volume with dH_2O. The reaction mix was incubated at 37°C for four hours. Digestion of the sample indicated the ancestral allele (T).

M9 (C→G): A 20 µl reaction was set up for each sample, containing 3 µl of amplified product, 2.5 units of *Hinf* I (New England Biolabs), 2 µl of 10x NE buffer 2 and made up to total volume with dH_2O. The reaction mix was incubated at 37°C for four hours. Digestion of the sample indicated the ancestral allele (C).

M175 (5bp deletion): A 20 µl reaction was set up for each sample, containing 3 µl of amplified product, 2.5 units of *Mbo* II (New England Biolabs), 2 µl of 10x NE buffer 2 and made up to total volume with dH_2O. The reaction mix was incubated at 37°C for four hours. Digestion of the sample indicated the ancestral allele (no deletion).

M122 (T→C): A 20 µl reaction was set up for each sample, containing 3 µl of amplified product, 5 units of *Nla* III (New England Biolabs), 2 µl of 10x BSA, 2 µl of 10x NE buffer 4 and made up to total volume with dH_2O. The reaction mix was incubated at 37°C for four hours. Digestion of the sample indicated the ancestral allele (T).

M17 (1bp deletion): A 20µl reaction was set up for each sample, containing 3 µl of amplified product, 5 units of *Afl*III (New England Biolabs), 2 µl of 10x BSA, 2 µl of 10x NE buffer 3 and made up to total volume with dH_2O. The reaction mix was incubated at 37°C for four hours. Digestion of the sample indicated the ancestral allele (no deletion).

M45 (G→A): A 20 µl reaction was set up for each sample containing 3µl of amplified product, 2.5 units of *Bfa* I (New England Biolabs), 2 µl of 10x NE buffer 4 and made up to total volume with dH_2O. Digestion of the sample indicated the ancestral allele (G).

Positive controls were used to ensure that the restriction enzyme digestion reactions were successful. This was achieved by including a previously genotyped sample in which successful digestion was known to have occured. As negative control, an undigested sample was loaded to aid in the identification of non-digested samples. To visualise the results of the digestion reaction, the entire 20 µl reaction mix was loaded on a 3% agarose gel and electrophoresed at 80V for 40 minutes

(agarose gel electrophoresis was as described in section 4.8). *Hpa* II digested pUC 19 DNA (Biotech) was also loaded as a size standard.

4.12 Mitochondrial DNA analysis

The mtDNA genome was studied for resolution of maternal ancestry. The study samples were tested for three mtDNA marker systems, namely the HV1 sequence in the non-coding D-loop, RFLPs in the coding region, and a 9bp indel in the COII/tRNA^{LYS} intergenic region.

Resolution of RFLPs proved problematic in this study due to the restricted amount of sample available. Therefore a three-stage process was adopted that optimised both sample use and haplogroup resolution. Firstly, the HV1 region was sequenced. Secondly, the presence or absence of the 9bp tandem repeats sequence (CCCCCTCTA) in the COII/tRNA^{LYS} intergenic region was detected in samples. The presence of the deletion indicated that the sample belonged to Haplogroup B. Thirdly, the remaining samples were then analysed for distinct HV1 motifs.

4.12.1 HV1 sequencing

DNA sequencing is based on the dideoxynucleotide (ddNTP) chain-termination method developed by Sanger *et al.* (1997). This method relies on the extension of a primer, hybridised to the template DNA to be sequenced, until a chain-terminating ddNTP is incorporated. There are four separate sequence reactions, each of which contains all four ddNTPs and a small proportion of one ddNTP. The ddNTPs lack the necessary 3'-OH for DNA chain extension, thereby terminating the growing oligonucleotide selectively at A, G, C, or T, depending on the dideoxynucleotide incorporated. As the incorporation is random, all possible truncated fragments will be produced, corresponding to the various positions of that particular base (Strachan and Read, 2004).

The first step in sequencing the HV1 region is to generate copies of the sequence through PCR. A 100 µl reaction was prepared for each DNA sample. This consisted of 5 µl of target DNA (5-10 ng/µl), 10 µl of 10x buffer solution (containing 1.5 mM $MgCl_2$ solution), 6 µl of 25 mM $MgCl_2$, 2.5 µl of 5 mM dNTPs, 2 µl of forward and reverse primer (100 ng/µl) (Appendix A4), 0.5 µl of Hot Star™ *Taq*

polymerase and 74 µl of dH_20. The thermal cycling conditions are listed in Table 4.7.

Table 4.7 mtDNA HV1 sequence amplification cycling conditions

No. of cycles	Temperature
1 cycle	94°C for 10 minutes
30 cycles	94°C for 45 seconds 66°C for 1 minute 72°C for 1 minute
1 cycle	72°C for 10 minutes

Following successful PCR amplification of the mtDNA HV-I region of each DNA sample, the PCR products were purified to remove all residual primers, nucleotides, polymerase and salts. The Qiagen QIAquick™ PCR purification kit was used because it allowed DNA fragments ranging from 100bp to 10,000bp to be purified. A 50 µl quantity of PB buffer was added to 10 µl of PCR product and the sample transferred into a QIAquick™ spin column (binding the DNA) in a collection tube. The column and collection tubes were centrifuged at 13,000 rpm for approximately 1 minute and the flow-through in the collection tube discarded.

The product remaining in the column (collection tube) was washed with 750 µl of PE buffer and centrifuged for one minute at 13,000 rpm. The flow-through was discarded and centrifuged again at 13,000 rpm for another one minute. The flow through and collection tube were discarded and replaced with a new collection tube. The DNA was eluted using 20 µl of dH_20, placed at the centre of the columnar membrane and allowed to stand for approximately one minute, then centrifuged at 13,000 rpm for one minute. The flow-through, which contained the purified PCR product, was kept and stored at –20°C.

The purified PCR product was sequenced using Applied Biosystems Ready Reaction Dye Primer™ (-21M13 and M13Rev) cycle sequencing kits. The sequencing kit contains four single pre-mix cycle sequencing reactions, i.e., one for each of the bases A, C, G and T. For bases A and C the reaction mixture consisted of 4 µl of pre-mix and 1 µl of purified PCR product. Bases G and T each consisted of 8

µl of pre-mix with 2 µl of purified PCR product. The sequencing reactions were placed in a thermal cycler under conditions listed in Table 4.8.

Table 4.8 mtDNA HV1 sequence reaction thermal cycling conditions

No. of Cycles	Temperature
15 cycles	94°C for 10 seconds 55°C for 5 seconds 70°C for 1 minute
15 cycles	96°C for 10 seconds 70°C for 1 minute
1 cycle	72°C for 10 minutes

Following the cycle sequencing reaction, the four extension reactions for each sample (i.e., A, C, G and T) were combined in a single microtube and mixed with 200 µl of 95% ethanol. The mixed extension products and ethanol were placed on ice for approximately 30 minutes to facilitate precipitation. Next, the samples were centrifuged at 13,000 rpm for approximately 30 minutes. The supernatant was removed and 250 µl of 70% ethanol were added to rinse and de-salt the sample. The resultant mixture was centrifuged at 13,000 rpm for 5 minutes, after which the supernatant was discarded and the DNA pellet was left at room temperature for approximately six hours until completely dry.

To prepare the pellet sample for loading into the polyacrylamide gel, 2.5 µl of loading buffer and 0.5 µl of deionised formamide was added to each pellet sample. The components and setup of the polyacrylamide gel are as previously described in section 4.8, 5 µl of each sample solution were pipetted into the gel wells and the gel was run for 16 hours using filter setting A. The resulting gel pictures were analysed with the Sequencescan™ software program. The Sequence Navigator™ version 1.01 software program (ABI Prism™) was then used to analyse the data extracted by the Sequencescan™ software program. For each sample, the forward sequence and the complementary reverse sequence were read in a parallel 5^1-3^1 direction and comparatively aligned. All nucleotide bases at either side that were not considered to be part of the HV-I region (e.g., primer sequence) were deleted.

The sequenced products were analysed using gel electrophoresis on the ABI 373A Automatic DNA Analyser (Applied Biosystems) with Sequence Navigator® software to edit the sequences. To resolve ambiguous bases and confirm sequence variants, PCR fragments were sequenced from both ends.

4.12.2 mtDNA 9-bp deletion detection

Detection of the presence or absence of the 9bp sequence (CCCCCTCTA) in the COII/tRNA^{LYS} intergenic region was accomplished by PCR of this fragment. A 10 µl reaction was prepared for each sample of DNA. This consisted of 5 µl of target DNA (5-10 ng/µl), 1 µl of 10x reaction buffer solution (containing 1.5 mM of $MgCl_2$), 1 µl of 25 mM $MgCl_2$ solution, 0.2 5µl of 5 mM dNTPs, 0.2 µl of forward and reverse primer (100 ng/µl) (see Appendix A5), 0.1 µl of Hot Star™ *Taq* polymerase, and 2.25 µl of dH_20. The thermal cycling conditions are listed in Table 4.9.

Table 4.9 COII/tRNA^{LYS} intergenic region thermal cycling conditions

No. of Cycles	Temperature
1 cycle	94°C for 5 minutes
30 Cycles	94°C for 15 seconds 45°C for 15 seconds 70°C for 30 minute
1 cycle	72°C for 10 minutes

The PCR product was resolved on a 4% agarose gel consisting of 3% NuSieve® fine grade agarose and 1% DNA grade agarose (Cambrex Life Sciences). Deletion detection was accomplished by identifying 100bp and 91bp bands on the gel. The presence of the smaller band indicates a deletion by reference to the size standard PUC19 (HpaII).

4.12.3 mtDNA haplogrouping protocol

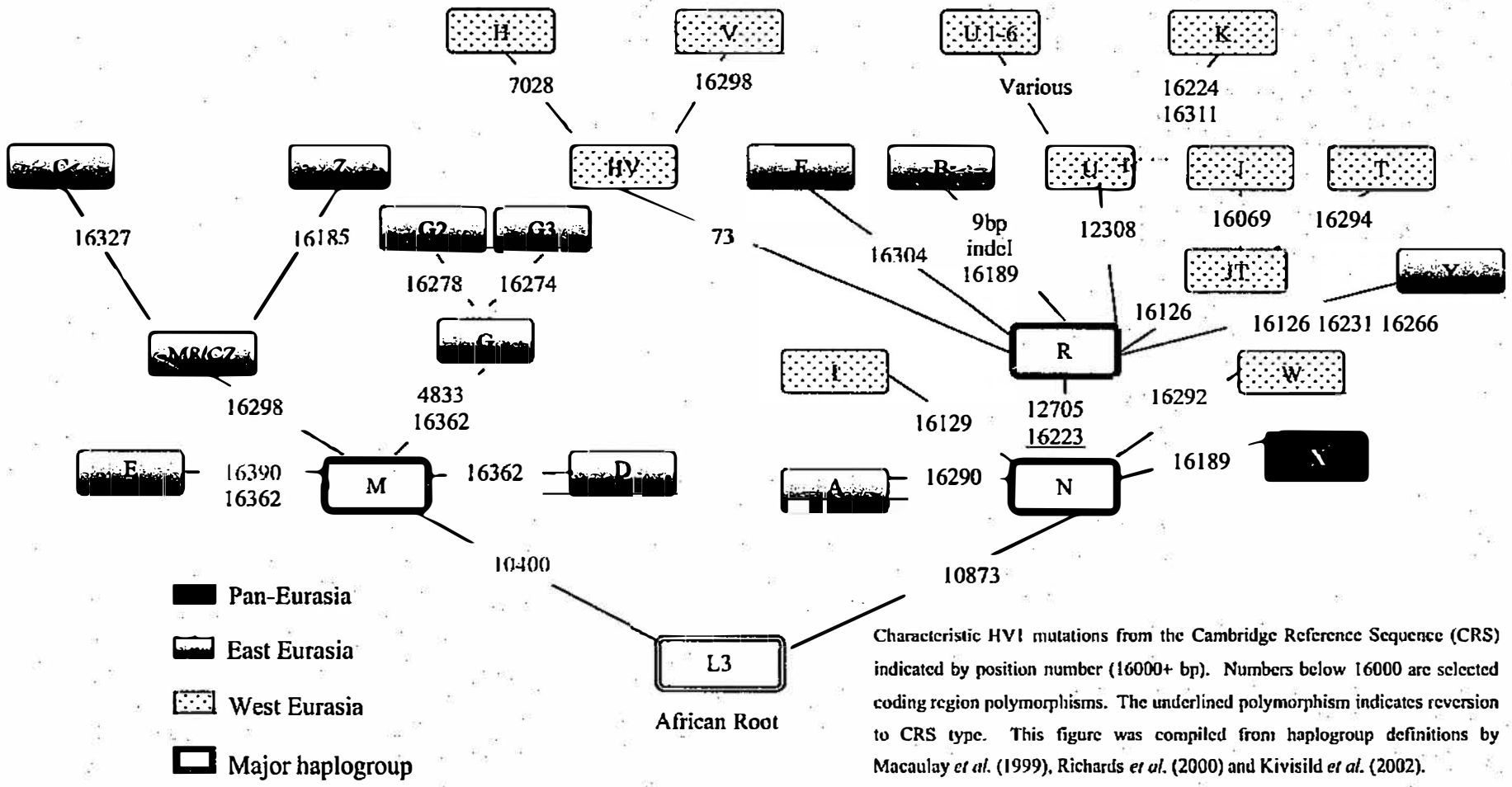
Methods for defining mtDNA phylogeographic haplogroups have been constantly evolving since the landmark studies by Torroni and co-workers (Torroni

et al., 1992; Torroni *et al.*, 1993; Torroni *et al.*, 1996). These haplogroups were defined by a combination of HV1 sequence motifs consisting of common mutations and RFLP polymorphisms in the coding region. More recent studies (Macaulay *et al.*, 1999; Kivisild *et al.*, 2002; Kivisild *et al.*, 2003) have described many haplogroups that are only defined with coding region polymorphisms, and they have recommended that haplogroups only be defined in this manner. However, other researchers consider that haplogroup identification can be achieved using HV1 motifs, using coding region definitions as a guide (Yao *et al.*, 2002a).

The degree of resolution of mtDNA haplogroups largely depends on the region of the world in which the study populations are resident. European (Macaulay *et al.*, 1999; Richards *et al.*, 2000), West Asian (Richards *et al.*, 2000; Quintana-Murci *et al.*, 2004), African (Salas *et al.*, 2002) and North Asian/American mtDNA haplogroups (Melton *et al.*, 2001) have been defined in great detail and resolution. Comprehensive investigations into haplogroup definition and diversity have commenced more recently in East Asia (Kivisild *et al.*, 2002; Yao and Zhang, 2002).

It is the haplogroup phylogenies developed from these investigations that are referred to in the current study. The haplogroups for this study are only broadly defined primarily with HV1 motifs, and with the 9bp deletion defining haplogroup B (Figure 4.5). This was found to be adequate for initial surveys of broad haplogroup identification and has been applied by Yao and Zhang, (2002) in defining haplogroup data they had gathered from previous studies.

Figure 4.5 Simplified Eurasian mtDNA HV1 haplogroup network



Characteristic HV1 mutations from the Cambridge Reference Sequence (CRS) indicated by position number (16000+ bp). Numbers below 16000 are selected coding region polymorphisms. The underlined polymorphism indicates reversion to CRS type. This figure was compiled from haplogroup definitions by Macaulay *et al.* (1999), Richards *et al.* (2000) and Kivisild *et al.* (2002).

CHAPTER 5

SUMMARY STATISTIC ANALYSIS

5.1 Introduction

The initial analysis of data from the study populations comprised a broad survey of summary statistic-based methods, in order to establish the nature of the genetic structure of each population and the direction to be taken for the subsequent analysis of intra- and inter-population relationships. The approach was to generate common population genetics summary statistics for autosomal and Y-chromosome microsatellites and mtDNA HV1 sequence data. This was followed by AMOVA and phylogenetic analysis to assess genetic structure and history.

As discussed in section 4.1, the populations were initially divided into two broad groups, 'Ancient' and 'Recent' based on historical evidence suggesting the relative ages of the populations. Thus for the 'Ancient' populations, their origins may be defined as far back as the Neolithic period i.e., 8000-10000 YBP (Du and Yip, 1993), while the formation of the 'Recent' populations can be dated back to 680AD (Gladney, 1998).

5.2 Autosomal microsatellite analysis

5.2.1 Autosomal summary statistics

The first part of the study was to review, update and reanalyse autosomal data from the nine study populations, two of which have been analysed in previous studies at the Centre for Human Genetics, ECU (Black *et al.*, 2001). It is important to state that DNA could not be extracted from some samples due to degradation, thus the blood spot sample numbers in Table 4.2 differ from the DNA samples stated in Table 5.1. In the selection of appropriate summary statistics, several factors were regarded as being of primary importance. Firstly, as these data were derived from random samples, no phase information was available and so there was no way of directly inferring the effect of recombination. For this reason, summary statistics based on general assumptions rather than particular mutational models were required, with statistics based on allele and genotype frequencies. The basic data generated from which the following summary statistics are derived are in Appendix B.3. Nei's gene diversity (Nei, 1987) was selected, as this statistic is sufficiently general in nature and also could be applied to Y-chromosome and mtDNA data, thus allowing comparison of population structure estimates between the different marker systems. Nei's gene diversity is defined as:

$$H = \frac{n(1-\sum x_i^2}{(n-1)}$$

Equation 5.1 Nei's Genetic Diversity.

where n is the number of gene copies and x_i is the frequency of the ith allele. The statistic measures the probability that two alleles drawn at random from a population will be different from each other and for diploid loci it is a measure of expected heterozygosity.

Gene diversity was compared to observed heterozygosity in the sample via the calculation of F_{IS} (Weir and Cockerham, 1984), which is the correlation measuring the probability of a match between two alleles from the same individual (I) compared with those drawn from the population (S) (Cockerham and Weir, 1973; Weir and Cockerham, 1984). If the population is divided into partially isolated subpopulations, individuals have an increased probability of sharing a common ancestor, and thus an increased probability of homozygosity (Overall and Nichols, 2001). A significant positive value indicates excess homozygosity, while a significant negative value indicates excess heterozygosity. To test significance the exact probability test was used, developed by Guo and Thompson (1992) from Fisher's exact test. The U-test developed by Rousset and Raymond (1995), which is an extension of the previous exact probability test, was further used to test the hypothesis of heterozygote excess or heterozygote deficiency in each population.

The results of the computation of these statistics from the autosomal data are shown in Table 5.1. A comparison of gene diversity and observed heterozygosity values shows some disparities. In theory, if a population is under Hardy-Weinberg equilibrium these values should be equal. However, as previously reported significant differences were exhibited for the Bo'an, Han, Hui and Sala (Black, 2000; Black *et al.*, 2001; Wang *et al.*, 2003). It was concluded that the low levels of heterozygosity, and hence the relatively large positive F_{IS} values, in the Bo'an, Hui and Sala could be attributed to their known history of population endogamy, whereas

Table 5.1 Summary statistics from autosomal microsatellite analysis

Population	*Sample size*	*Gene diversity*	*Observed heterozygosity*	*Average no. of alleles per loci*	F_{IS}	*HWE exact test p value*	*Heterozygote deficiency U Test p Value*	*Heterozygote Excess U Test p Value*
				"Ancient" Populations				
Han	94	0.7190	0.6391	10.8	0.1132	~0	~0	~1
Kucong	39	0.6880	0.7489	7.7	-0.0899	0.069	0.996	0.004
Miao	42	0.7380	0.7766	9.2	-0.0539	0.016	0.877	0.095
Tibetan	22	0.6917	0.6855	7.7	0.0040	0.02	0.117	0.806
Yao	14	0.6179	0.6107	5.4	0.0048	0.8	0.253	0.7575
				"Recent Populations"				
Bo'an	67	0.7786	0.6739	12.6	0.1360	~0	~0	~1
Dongxiang	55	0.7368	0.7276	10.6	0.0151	~0	~0	~1
Hui	53	0.7504	0.4521	9.8	0.4021	~0	~0	~1
Sala	76	0.7746	0.6618	11.8	0.1443	~0	~0	~1

for the Han the low observed heterozygosity was more likely to result from unrecognised population substructure (Du, 1981; Black *et al.*, 2001; Wang *et al.*, 2001; Wang *et al.*, 2003). The low F_{IS} value in the Dongxiang could be attributed to their large population size by comparison with the Sala and more especially the Bo'an. Nationwide the Hui have a population of approximately 9.8 million, but they consist of many geographically isolated communities each having a different history and therefore different, possibly restricted, population structures (Lipman, 1997; Gladney, 1998).

A second factor of importance in interpreting genetic structure is the average number of alleles per locus in each population. By comparing these figures with F_{IS} values, it is apparent that populations with a higher average number of alleles per locus also have positive F_{IS} values. Therefore, even though there is a larger pool of alleles in these populations, a significant level of homozygosity also exists. This suggests that there may be population sub-structuring within the Bo'an, Sala, Hui and Han, with specific alleles restricted to certain groups in the population. For the Muslim populations, it provides additional evidence to support the proposal that sub-community endogamy and consanguinity have exerted significant effects on their genetic structure, as these practices would restrict the flow of alleles within the population as well as with other co-resident populations.

The samples from the Miao, Tibetans and Yao have not been previously analysed or published and, overall, they show no evidence of significant heterozygote deficiency or excess. Each of these populations have had much longer histories, with significantly more generations during which genetic drift could operate. In addition, Tibetan communities strenuously avoid consanguineous marriage and the effect of this avoidance may also be to maintain heterozygosity.

The Kucong show evidence of a very different structure. While the exact test did not detect any statistically significant deviation from HWE at the 95% confidence level, application of the more specific U test for heterozygote excess attained statistical significance ($p = 0.004$). Therefore the negative F_{IS} value in the Kucong is significant and indicates heterozygote excess in the population. In human populations, heterozygosity of this type is often explained as the result of either heterozygote advantage or avoidance of consanguinity. It is hard to judge which

effect, if not both, is behind the excess of heterozygosity in the Kucong due a near-complete lack of information on their history and demography.

5.2.2 Autosomal AMOVA analysis

A series of AMOVA analyses were undertaken to compare relationships between the study populations in regard to their autosomal structure. Five different grouping structures were tested, two based on geographical location, two on cultural/historical grounds and one on population structural grounds according to the results from the summary statistics (Table 5.1). The first geographical structure tested split the populations into groups of North East, Central and Southern China, while the second geographical grouping scheme included Tibetans as a fourth group representing the South-West of the country.

The 'historical' grouping schemes were based both on historical information and previous genetic studies. Thus historical grouping (a) first split the populations into three groups, the northern Han, Muslim populations, and southern non-Han populations. The second grouping placed the Tibetans into a separate fourth group according to studies that have indicated they are of northwestern rather than southern genetic origins (Torroni *et al.*, 1994; Yao *et al.*, 2002b; Wen *et al.*, 2004b). The fifth grouping scheme represented differences in the observed internal population genetic structure of the study populations, which were variously grouped according to heterozygosity deficiency, no deviation from HWE, and heterozygote excess respectively (Table 5.2).

Two methodologies pertain when conducting AMOVA on microsatellite data, the difference in the number of alleles and the sum of squared allele size difference. According to Mickalakis and Excoffier (1996), the first method is analogous to F_{ST} analysis (Wright 1952, Weir & Cockerham 1984), and the second is analogous to R_{ST} (Slatkin, 1995). The second method was chosen as it was not affected by sample size variation and was less affected by missing data, and so was statistically more robust (Goldstein *et al.*, 1995; Goldstein *et al.*, 1996). This methodology was utilised for both autosomal and Y-chromosome microsatellites.

The results of the AMOVA analysis are shown in Table 5.3 and show that, whichever model was used, the proportion of variance among populations was similar, ranging from 4.37% to 7.35%. Only models geographical (b) (6.44%) and

Table 5.2 Inferred Group Structures for AMOVA analysis

	Group 1	Group 2	Group 3	Group 4
Null	All populations	N/A	N/A	N/A
Geographic(a)	Han, Hui	Bo'an, Dongxiang, Sala	Kucong, Miao, Tibetan, Yao	N/A
Geographic(b)	Han,Hui	Bo'an, Dongxiang, Sala	Kucong, Miao, Yao	Tibetan
Historic(a)	Han	Bo'an, Dongxiang, Hui, Sala	Kucong, Miao, Tibetan, Yao	
Historic(b)	Han	Bo'an, Dongxiang, Hui, Sala	Kucong, Miao, Yao	Tibetan
Structural	Han,Bo'an,Hui,Sala	Dongxiang. Miao, Yao, Tibetan	Kucong	

N/A = not applicable

Table 5.3 AMOVA analysis of autosomal genotypes

Grouping scheme	No. of groups	Variance components (%)		
		Within populations	Among populations within groups	Among groups
Null structure	1	94.22*	5.78*	
Geographical(a)	3	93.80*	4.37*	1.83
Geographical(b)	4	92.45*	6.44*	1.22
Historical(a)	3	94.12*	5.52*	0.36
Historical(b)	4	91.41*	4.71*	3.88*
Structural	3	92.00*	7.35*	0.66

* statistically significant at a 95% confidence limit.

structural (7.35%) displayed proportions of variance among populations larger than the assumption of no hierarchical structure (5.78%). However, the historical (b) model displayed a genetic variance among groups (3.88%) that was statistically significant at a 95% confidence level. Therefore, under this model, the total proportion of variance other than within-populations is 8.59%, which is greater than the result for null structure.

The structural and historical (b) models showed the greatest proportion of genetic variance not within populations. While both results infer the major role that internal structure exerts on variance between the study populations, the significant result for the historical (b) grouping also suggests that ancestral origins have a role to play. In any case, the results support the idea that differences between the populations are affected by two major factors, founder effect and, to a larger extent, the subsequent action of genetic drift.

5.2.3 Autosomal phylogenetic analysis

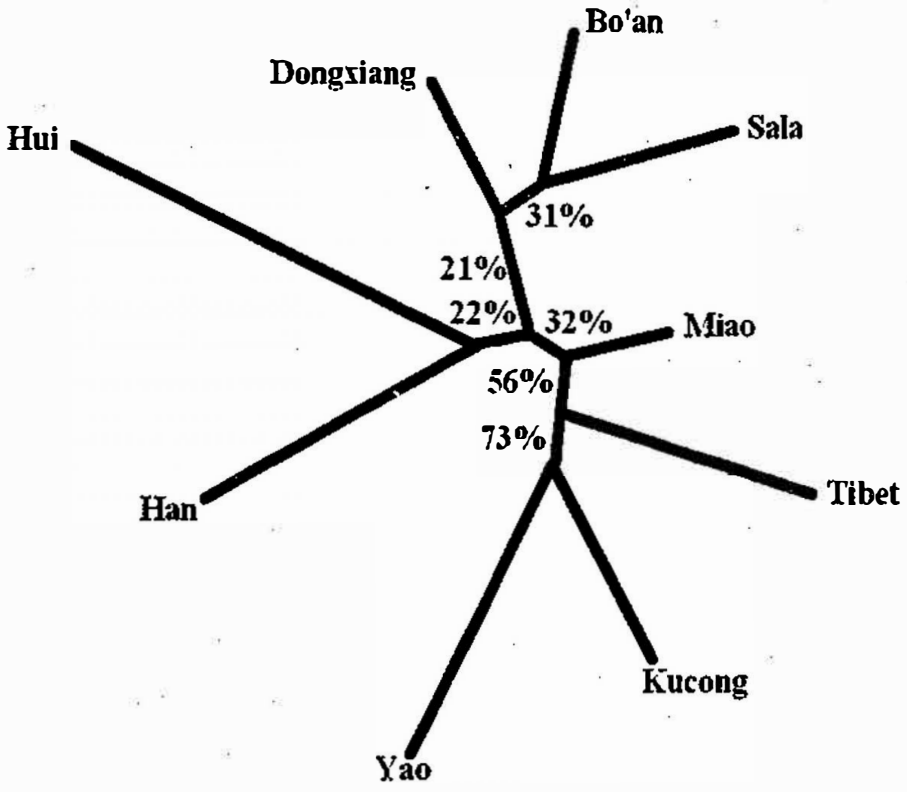
The final stage in this initial analysis of autosomal data was to calculate genetic distances between the study populations. The genetic distance chosen for autosomal microsatellites was Nei's Standard distance, a distance based on allele frequency. This more general distance was chosen because, as previously indicated, phase information was not available and so a distance measure based on a particular mutational model was not appropriate.

Using the software MICROSAT, a consensus matrix based on bootstrapping 1000 samplings was produced. From the consensus distance matrix an unrooted, neighbour-joining tree was plotted (Figure 5.1). As trees give the impression of direct population splitting, rather than genetic similarity/dissimilarity, multidimensional scaling analysis was then performed on the consensus matrix data and a 2D scatter plot produced from the resulting data (Figure 5.2).

The Neighbour-joining tree indicates clustering of the study populations, with the Bo'an, Dongxiang and Sala in one cluster and the Kucong, Miao, Tibetans and Yao in another cluster (Figure 5.1). The Han and the Hui seem to cluster together. However, as the branch lengths between them are quite long, this impression is

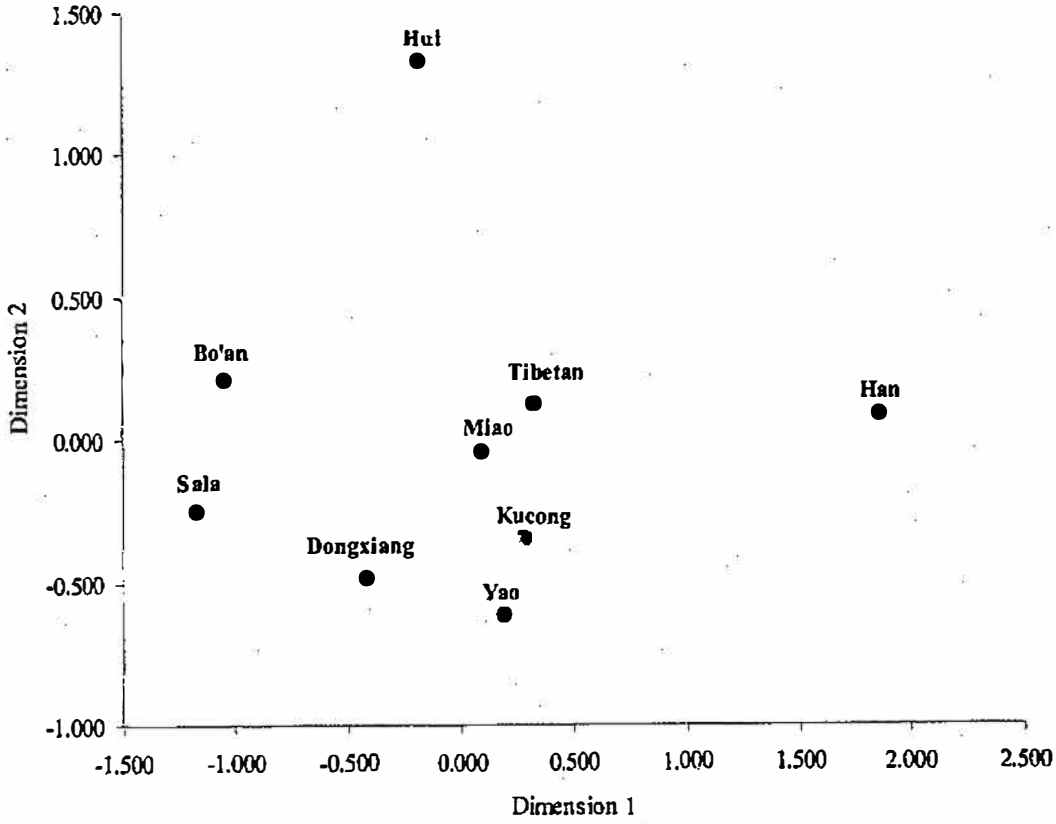
deceptive. This feature is clearly shown in the MDA plot, Figure 5.2, where the Han and Hui are clearly separated while the other populations are more closely bunched.

Figure 5.1 Consensus unrooted Neighbour Joining Tree constructed from genetic distances calculated for autosomal microsatellite genotype data from the nine study populations



*Percentages represent the proportion of bootstrap samples containing that branch.

Figure 5.2 Multidimensional Scaling Analysis based on genetic distances calculated for autosomal microsatellite genotype data from the nine study populations



The most striking aspect of the MDA plot is the large difference between the Hui and the other populations, with their higher level of homozygosity a possible explanation. The Han are also distinct from the other populations, although less so than the Hui, and once again their significant level of homozygosity may be an important consideration. The Bo'an and Sala are grouped together and both populations had significant positive F_{IS} values. Finally, the populations that had no significant F_{IS} values are clustered. On these bases it can be argued that the distances calculated from the relatively small number of autosomal loci examined are indicative of differences in internal population genetic structure.

In summary, this initial analysis of the autosomal data shows that the genetic differences between the study populations are probably less dependant on their ancestral origins than on the effects of genetic drift following their foundation. Subsequent admixture in the 'Ancient' populations, and endogamy and consanguinity in the 'Recent' populations, are additional contributory factors. These different influences are further explored using likelihood-based modelling techniques in Chapter 7.

5.3 Y-chromosome microsatellite analysis

5.3.1 Y-chromosome microsatellite summary statistics

The next stage in the study was to analyse the paternal genetic histories of male samples from the previous autosomal analysis via Y-chromosome microsatellite analysis. During the collection of the data some of the Y-chromosome microsatellite markers tested could not be resolved for all populations. This was due to a range of factors, including the limited amount of DNA available from the blood spot samples and the age and quality of these samples. Some markers could not be resolved in specific populations due to an adequate sample size. For example, the marker DYS392 could not be adequately resolved in the Sala, although this was possible for the other populations. In practice, for adequate comparisons of all nine populations, a haplotype composed of markers DYS19, DYS388, DYS389a, DYS389b and DYS393 was the best possible outcome. As will be demonstrated, this short haplotype was informative for both intra- and inter-population diversity analysis. The basal experimental data are contained in Appendix B.2 and the summary statistics generated from them in Table 5.4.

Table 5.4 Summary Statistics from Y-chromosome microsatellite analysis

Population	Sample size	No. loci	Gene diversity	Average no. alleles per locus	Haplotypic diversity	θ_k
		'Ancient' populations				
Han	91	5	0.652	5.6	0.995	186.17
Kucong	21	5	0.673	3.8	0.986	91.56
Miao	32	5	0.592	4.0	0.978	41.95
Tibetan	20	5	0.565	3.2	0.963	25.59
Yao	13	5	0.590	3.4	0.961	18.02
		'Recent' populations				
Bo'an	40	5	0.579	3.8	0.990	72.11
Dongxiang	41	5	0.599	4.2	0.991	110.37
Hui	22	5	0.671	4.0	0.987	63.01
Sala	32	5	0.593	4.4	0.994	72.11

Analysis of the Y-chromosome data displayed some interesting findings. For example, the marker DYS388 showed wide diversity in the study populations exhibiting eight different alleles (Table 5.5). This is remarkable as DYS388 has not been commonly used in population studies because it had been reported to have low allele diversity (Kayser *et al.*, 1997). For this reason it was not included in the original forensic marker panel constructed by Kayser *et al.*, (1997) or in the Y-chromosome Haplotype Reference Database (http://yhrd.org). While low allele diversity has since been confirmed in European populations (Dupuy *et al.*, 2004), higher allele diversity has been found in populations outside Europe, including North Africa (Bosch *et al.*, 2000), the Middle East (Nebel *et al.*, 2001) and East Asia (Kwak *et al.*, 2005). In the present study, DYS388 showed largest diversity in the Dongxiang, Han, Hui and Sala, who contained six, seven, six and six alleles respectively (Table 5.5). As multiple ancestries have been claimed for these populations, it is hypothesised that their relatively high diversity for DYS388 may be indicative of their diverse founder origins.

The basic summary statistics (Table 5.4) did not indicate any significant between-population differences. Gene diversity values ranged from 0.565 in the Tibetans to 0.673 for the Kucong, with an average of 0.613. By referring to the average number of alleles one can surmise that the gene diversity figure for each population is the result of different factors. For example, the Han had a gene diversity value of 0.652 and the Kucong 0.673. However, they have mean allele numbers of 5.6 and 3.8 respectively. The fact that the Kucong had a slightly higher gene diversity value, but a lower average allele number amounting to almost one allele, indicates a possible gap in the distribution of repeats in the Kucong. This may indicate differential reproductive opportunities within each population, with the Kucong having a more limited paternal gene pool than the larger Han population.

A further indication of the available paternal gene pool in each population can be inferred from estimation of the key population genetic parameter θ, termed the effective genetic diversity (Jobling *et al.*, 2004). The parameter θ represents the product of effective population size and the mutation rate. This parameter will be the subject of further analysis via summary statistical

Table 5.5 DYS388 allele frequencies in the nine study populations

Allele length (repeats)	Allele Frequency (%)								
	Bo'an	Dongxiang	Hui	Sala	Han	Kucong	Miao	Yao	Tibetan
10	2.50	7.32	-	3.13	17.58	-	18.75	-	-
11	-	2.44	-	3.13	2.20	19.05	-	7.69	15.00
12	70.00	60.98	54.55	68.75	52.75	33.33	75.00	23.08	70.00
13	12.50	19.51	9.09	3.13	16.48	38.10	3.13	7.69	0.00
14	15.00	2.44	18.18	12.50	5.49	9.52	3.13	61.54	15.00
15	-	7.32	9.09	9.38	2.20	-	-	-	-
16	-	-	4.55	-	-	-	-	-	-
17	-	-	4.55	-	3.30	-	-	-	-

methods in the haploid systems only. The effective genetic diversity will be inferred for autosomal data with the use of likelihood-based modelling in which recombination can be addressed. The estimation of θ permits a measure of effective population size, N_e, a parameter that is indicative of the size of the available paternal gene pool.

For this summary statistic section, an estimate of θ termed θ_K was calculated (Ewens, 1972). This estimate is taken from the calculation of estimated number of alleles and involves an infinite-allele equilibrium relationship between the expectation (k), the sample size (n) and θ where:

$$E(k) = \theta \sum_{i=0}^{n-1} \frac{1}{\theta + i}$$

Equation 5.2 Estimated Number of Alleles

This method of θ estimation was used by (Zerjal *et al.*, 2002), where it was demonstrated that the value of θ was an indicator of the existence of population-specific clusters of haplotypes, rather than dispersed haplotypes from multiple ancestries. However, as the estimate is related to sample size, sample size bias must be considered.

Most populations exhibited relatively high θ_K values, indicative of multiple male ancestries. This included all the Muslim populations, as well as the Han population for which historical sources have indicated that multiple ancestries are likely. The Kucong also exhibit a high θ_K value (91.56) thus suggesting a multiple male ancestral structure, although this value is less than half that estimated for the Han (186.17). The result fits with the earlier hypothesis comparing the gene diversity value and mean number of alleles. That is, while diverse the Kucong appear to have a much more restricted available paternal gene pool than the larger Han population. The other three southern Chinese populations, the Miao, Tibetans and Yao, exhibit lower θ values, which suggests the possibility of restricted male ancestries. Caution must, however, be exercised in drawing conclusions from the results for the Yao due to their small sample size, which in turn probably influences

the θ value. Nevertheless, the Miao have a bigger sample size than the Kucong and yet they had a θ value less than half as large (41.95 versus 91.56). This would appear to indicate that the Miao have a restricted paternal gene pool and thus a population-specific haplotype cluster, but the wide variance in θ values may also be due to the wide divergence in sample sizes, from 91 Han to just 13 Yao samples.

5.3.2 Haplotype analysis

The possibility of population-specific haplotype clusters and the variation in sample sizes led to direct haplotypic analysis. From the list of haplotypes presented in Appendix B.2, a summary of haplotype sharing was compiled (Table 5.6). These summary data show that each population has a large proportion of population-specific haplotypes. This is a striking result since only five markers were adequately resolved in all nine populations. A large range of haplotypes were found in the study populations indicative of diverse paternal ancestries, and especially so in the 'Recent' populations, although also evident in the Han and in the Tibetans.

The Kucong, Yao and Miao are largely composed of unique haplotypes with the Kucong haplotypes mainly restricted to the small 13 repeat DYS19 allele. This can be inferred as being an older haplotype than the other populations. The mutational behaviours of these microsatellites, especially DYS19, have been widely studied and it has been concluded that there is a tendency for mutations to occur in a cumulative fashion, with the number of repeats increasing rather than decreasing through time (Goldstein and Pollock, 1997; Goldstein and Chikhi, 2002).

The raw haplotype results could be biased due to the variable sample sizes for each population, ranging from 79 Han haplotypes to just 13 Yao haplotypes. Even so, the Tibetan, Yao and Miao haplotypes showed a wide range of allele sizes, most of which were unique to each population.

One can conclude that the large numbers of unique haplotypes in each population could be the result of restricted paternal migration. This has been reported in historical sources which have indicated that, in China, it is more likely that females, rather than males, migrated and assimilated into other communities due to the largely patriarchal societal structure which by its nature, is composed of a restricted number of male lineages (Du and Yip, 1993; Yao *et al.*, 2000; Yao *et al.*, 2002b; Yao and Zhang, 2002).

Table 5.6 Summary of haplotype-sharing between the study populations*

Population	No. of haplotypes	Population-specific haplotypes	Shared haplotypes
	'Ancient' populations		
Han	79	54	25
Kucong	21	17	4
Miao	32	23	9
Tibet	20	14	6
Yao	13	12	1
	'Recent' populations		
Bo'an	40	17	23
Dongxiang	41	22	19
Hui	22	12	10
Sala	32	25	7

* The data in this table refers to comparison of haplotypes between populations, regardless of haplotype sharing within a population.

5.3.3 Y-chromosome microsatellite AMOVA analysis

The paternal genetic structure of the study populations was statistically assessed by AMOVA analysis in a similar manner to the autosomal data, with the populations were grouped using the same scheme as in section 5.2.2. As for the autosomal data, a SMM model of mutation also was assumed.

The results shown in Table 5.7 indicate that the geographical and historical grouping models displayed aberrant results and, with a large negative proportion of variance values recorded, these grouping schemes are not valid for the analysis of the study populations on a paternal basis.

What is also apparent is that the proportion of variance among populations for each structure was almost twice as large as the level recorded in the autosomal gene pool. This observation indicates the existence of diverse paternal histories for each population, which at least in part was responsible for the previously elucidated unique haplotypic distributions in each population, and hence strong differentiation between them.

Table 5.7 AMOVA analysis of Y chromosome microsatellite diversity in the 9 study populations

Grouping scheme	No. of groups	Variance components (%)		
		Within-populations	Among - populations within groups	Among groups
Null structure	1	83.78*	16.22*	
Geographical (a)	3	84.65*	19.94*	-4.59
Geographical (b)	4	85.06*	22.95*	-8.01
Historical (a)	3	85.08*	21.26*	-6.34
Historical (b)	4	85.59*	24.7*	-10.29
Structural	3	82.77*	13.98*	3.25*

* statistically significant within a 95% confidence interval.

5.3.4 Y-chromosome phylogenetic analysis

The investigation of differing paternal histories in each population was further explored via phylogenetic analysis, which involved generation of a consensus neighbour-joining tree from bootstrap sampling of all 126 possible samples.

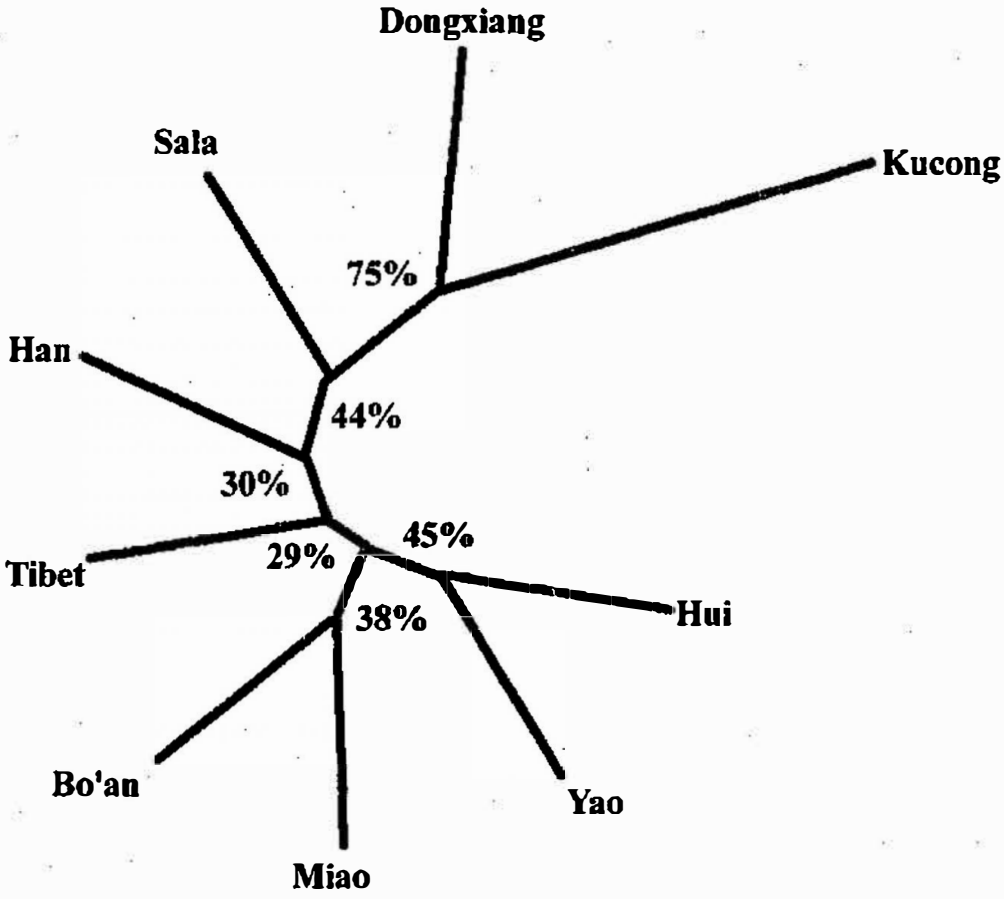
Figure 5.3 shows that there is no immediately obvious grouping of populations equivalent to that found for the autosomal data. Examples of seemingly incongruous relationships are the Hui grouping closely with the Yao, and the Bo'an and Miao co-clustering. On initial inspection such groupings could infer close ancestral ties, but the histories of these pairs of populations are quite different (Du and Yip, 1993). Therefore it seems more likely that the clustering displayed by this tree links populations by similar internal structuring rather than on an ancestral basis.

With further investigation of the consensus distance matrix by MDS analysis (Figure 5.4), the resulting scatter plot displays disparate genetic distances between the populations. By reference to historical sources (Du and Yip, 1993; Lipman, 1997; Gladney, 1998), it can be inferred that the clustering pattern displays no relationships based on the previously assumed geographical or historical similarities

between the populations. As the results are based on only five microsatellite markers, it is however concluded that there is insufficient resolution from such a small haplotype to infer any ancestral relationships and thus the relationships are purely structural. This is certainly the case if the results are considered in tandem with the AMOVA analysis, showing significant differences in variance according to structural differences.

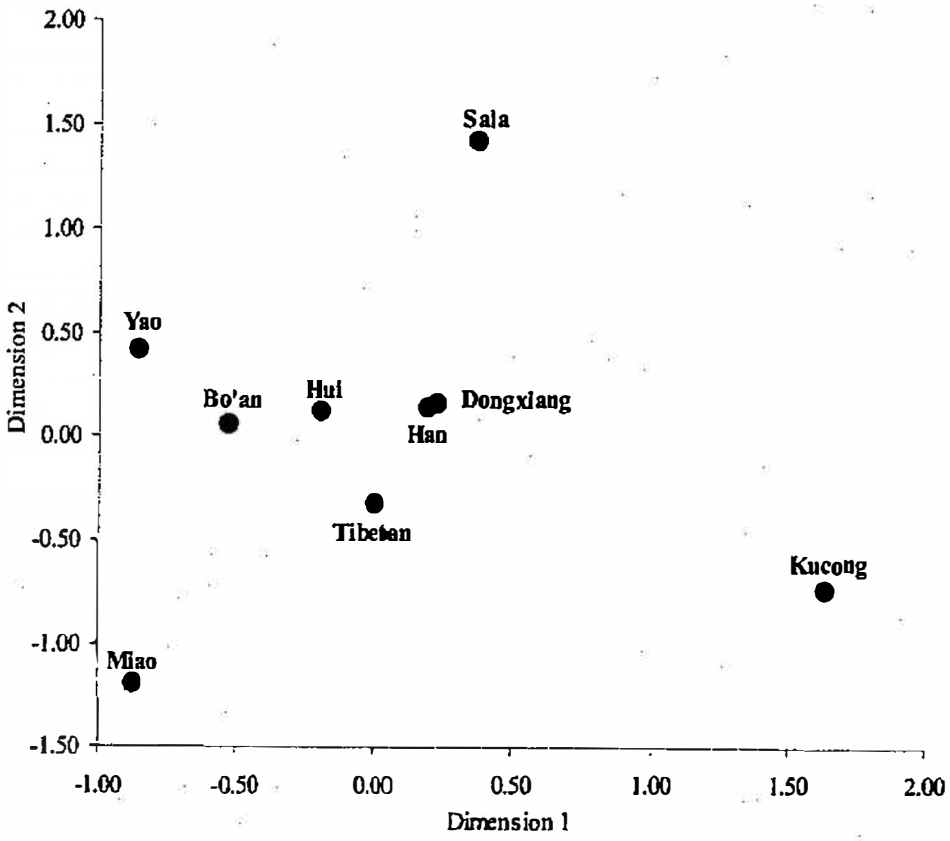
In figure 5.4, the Kucong are significantly different from all other populations and function as an outlier in the plot. By reference to the NJ tree only, one could be misled into inferring a relationship between the Kucong and Dongxiang. However, the branch length between them is large and this difference is more clearly demonstrated on the MDS plot. To further explore the origins of these differences, ancestral inference via UEP analysis is required. This possibility is explored in Chapter 6.

Figure 5.3 Unrooted Neighbour Joining Tree based on genetic distances calculated from Y-chromosome microsatellite haplotypes of the nine study populations



*Percentages represent the proportion of bootstrap samples containing that branch.

Figure 5.4 Multidimensional Scaling Analysis based on genetic distances calculated for Y-chromosome microsatellite haplotypes from the nine study populations



5.4 Mitochondrial HV1 sequence analysis

5.4.1 mtDNA summary statistics

As a significant differentiation of the populations could be ascertained on a paternal basis, the next step was to investigate the possibility of comparable differentiation in terms of maternal diversity. For this purpose mitochondrial HV1 sequences were resolved for all study populations. While the sample sizes were small the calculation of summary statistics nonetheless revealed interesting aspects of the maternal genetic structure of the study populations. The results are shown in Table 5.8 and the basal data they were derived from are presented in Appendix B.1.

The number of segregating sites (S) is simply a count of the number of nucleotide sites that vary over the entire set of aligned sequences, and they ranged from 39 in the Bo'an to 25 for the Sala. Simply counting the number of segregating sites cannot, however, provide a true estimate of the diversity in each population since it is dependant on sample size. Therefore a statistic analogous to gene diversity, termed estimated mean nucleotide diversity, was calculated as shown in Equation 5.3.

$$\pi = \frac{n(\sum x_i x_j \pi_{ij})}{n-1}$$

Equation 5.3 Mean Nucleotide Diversity

where n is the number of sequences, x_i and x_j the frequencies of ith and jth sequences respectively, and π_{ij} is the proportion of different nucleotides between them. As shown in Table 5.8, the estimation of nucleotide diversity in each population showed similar values, ranging from 0.018 for Tibetans to 0.025 for the Han, thus indicating little between-population differentiation.

The next stage in the analysis was to assess maternal demographic structure, which initially was achieved by calculating Tajima's D test and Fu's F_S statistic. While both measures primarily test statistics for the assumption of selective neutrality, they also are useful as evidence of demographic growth and thus present evidence of population size variation. Tajima's D test (Tajima, 1989a, 1989b) compares two estimators of the effective genetic diversity parameter θ, one based on the number of segregating sites in the sample, and the other on the mean number of

Table 5.8 Summary statistics from mtDNA HV1 sequence analysis

	n	*S*	π	Tajima's *D*	*D* p value	Fu's *Fs*	*Fs* p value	τ	*r*	Expansion Time (YBP)
				'Ancient' populations						
Han	11	35	0.025	-1.293	0.104	-5.169	0.006	7.002	0.052	59,000
Kucong	11	31	0.024	-0.862	0.219	-4.712	0.016	7.207	0.027	61,000
Miao	9	30	0.023	-1.029	0.175	-3.784	0.015	6.135	0.046	52,000
Tibetan	15	33	0.018	-1.178	0.128	-10.451	~0	4.541	0.035	38,000
Yao	10	31	0.020	-1.294	0.103	-1.294	0.001	5.659	0.037	48,000
				'Recent' populations						
Bo'an	14	39	0.023	-1.287	0.104	-9.306	~0	5.233	0.025	44,000
Dongxiang	9	29	0.022	-1.103	0.154	-2.693	0.056	6.586	0.049	55,000
Hui	13	29	0.020	-1.181	0.129	-6.945	0.001	7.253	0.027	61,000
Sala	11	25	0.019	-0.875	0.216	-5.570	0.003	5.052	0.045	42,500

107

pairwise differences between haplotypes. Under the infinite site model, both estimators should measure the same quantity, but differences can arise under selection, in non-stationary populations, or because of heterogeneity in mutation rates among different sites.

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_S}{\sqrt{Var(\hat{\theta}_\pi - \hat{\theta}_S)}}$$

Equation 5.4 Tajima's D statistic

Where $\hat{\theta}_\pi = \hat{\pi}$ = mean number of pairwise differences, $\hat{\theta}_s = S/\sum_{i=0}^{n-1}(1/i)$, and S is the number of segregating sites in the sample.

Fu's F_S statistic (Fu, 1997) tests the assumption of neutrality in a different manner. While it also assumes the infinite site model, it is based on the probability of observing N or more alleles in a sample of a given size, conditioned on the observed average number of pairwise differences. It is calculated by evaluating the probability of observing a random neutral sample with a number of alleles similar to, or smaller than, the observed value given the observed number of pairwise differences, which is taken as an estimator of θ. Formally, this probability is called S', where $S' = P(K \geq k_{obs} \mid \theta = \hat{\theta}_\pi)$ and Fu's statistic is defined as the logit of S' (Fu, 1997).

Fu (1997) reported that the F_S statistic was very sensitive to population demographic expansion, which generally leads to large negative *Fs* values. In addition, the probability distribution has its 5% cutoff value, i.e. a 5% significance interval at the second percentile, and so the statistic is only significant at a p value 0.02.

For all populations a non-significant negative D value was found, which is believed to signify past population expansion. The more sensitive *Fs* test shows that eight of the nine populations had statistically significant negative values, all of which were relatively large. This indicates that the populations are undergoing expanding population growth. However, as the Muslim populations are relatively young in historical terms, growth may not be the major factor. It seems more likely that the F_S results are due to the maternal ancestors being mostly of Han origins and thus

exhibiting an older history than the populations themselves. In the case of the Miao, Tibetan and Yao minority populations a history of continuous migration is probably the main explanatory factor for the F_S findings, as evidenced in previous genetic studies and from historical evidence (Du and Yip, 1993; Su *et al.*, 1999). In the case of the Dongxiang this is not evident perhaps, as explored in the autosomal analysis, to differences in population size. However, additional investigation would be required to confirm or reject this assumption.

5.4.2 mtDNA mismatch distributions

Mismatch distributions were computed and the relevant τ and r statistics were calculated to further explore this aspect of maternal demographic structure. The mismatch distribution (Rogers and Harpending, 1992) is based on the distribution of pairwise differences and is commonly used in the analysis of mitochondrial sequences. In effect, the mismatch distribution summarizes discernible genetic diversity, but in addition, the shape of the distribution has been shown to indicate aspects of population history, mainly episodes of population expansion. The indication of shape is given by the r statistic. A mismatch distribution with an r statistic less than 0.03 is considered smooth, i.e., unimodal and indicative of past population expansion, much like that indicated by the D and F_S statistics. As shown in Table 5.8, the r statistics for all nine populations averaged 0.038, with the Bo'an, Hui and Kucong exhibiting values below 0.03. But as the sample sizes are small, the other values for the six other populations could also be considered to indicate unimodality.

The advantage offered by mismatch distribution analysis is that one can infer the τ parameter, growth or decline measured in units of mutational time and defined as $\tau = 2ut$, where u is the mutation rate per sequence per generation and t is time in years. Values for τ ranged from 4.541 for the Tibetan population to 7.253 for the Hui. Assuming a divergence rate of 0.33% per million years (Ward *et al.*, 1991; Yao *et al.*, 2002b), this results in t values ranging from ~38,000 years to ~61,000 years, which correspond with the arrival of modern humans in East Asia. This is not to say that any of these populations were in existence near to this time period, however it serves as an indication that all of the populations probably have some maternal ancestral origins based in East Asia.

5.4.3 Maternal genetic structure

An AMOVA analysis of the mtDNA data was completed and compared with the autosomal and Y-chromosome results. The measure used was based on the Tamura Nei model of nucleotide mutation with a gamma correction of 0.3. The results shown in Table 5.9 indicate that AMOVA analysis showed very little hierarchical structure in the mtDNA data, lending further evidence to the belief that the maternal gene pools of the study populations were homogeneous. Only a small proportion of variance among populations was resolved, demonstrating the slight difference between the populations. This difference is most likely to be due to geographical factors, as geographical (b) showed the largest significant proportion of variation due to amongst group differences (4.44%). This is very different to the Y-chromosome AMOVA results (Table 5.4), where it was the structural grouping scheme which displayed greatest among-group variance. As the sample size for the mtDNA analysis was much lower than for the Y-chromosome analysis, the significance of this comparison is, however, questionable.

What can definitely be concluded from the mtDNA data is that the 4-5% of total variance that could be apportioned to among-population or among-group variance is much lower than the equivalent value of 16-17% in the paternal gene pool. Therefore it can be inferred that while they have widely different paternal gene pools, the study populations have a largely shared maternal history.

5.4.4 mtDNA phylogenetic analysis

Further exploration of the possibility of shared maternal histories was undertaken by phylogenetic analysis. For this purpose the Tamura-Nei distance with gamma correction was chosen, as discussed in section 3.4.1 (Nei and Kumar, 2000). From the matrix of distances calculated, an unrooted neighbour-joining tree was constructed and MDS analysis performed. Unlike the autosomal and Y-chromosome tree, the mtDNA tree was constructed from the distance mean between population distances calculated in the MEGA program. Therefore the tree is constructed from mean distance values, not as a consensus from multiple matrices, and hence no bootstrapping values are shown. The results of both analyses indicate a large degree of homogeneity between eight of the nine populations. However phylogenetic analysis did reveal differences between the Kucong and the other study populations.

Table 5.9 AMOVA analysis of mtDNA HV1 sequence diversity in the nine study populations

Grouping scheme	No. of groups	Variance components (%)		
		Within populations	Among populations, within groups	Among groups
Null structure	1	95.98*	4.02*	-
Geographical (a)	3	95.73*	3.59*	0.68
Geographical (b)	4	95.20*	0.36	4.44*
Historical (a)	3	96.22*	4.77*	0.99
Historical (b)	4	95.32*	1.78	2.90*
Structural	3	95.32*	2.75*	1.93

- statistically significant at a 95% confidence level.

The NJ tree was a classic star-shape and, according to established population genetics theory, a star-shape phylogeny indicates rapid population growth and/or homogeneity. In this case, it is almost certainly due to the small distances and thus the large-scale homogeneity among the study populations, as evidenced by the comparative scale in Figure 5.5 of just 0.002 distance units. The Kucong have a larger branch than the other populations on the tree signifying a maternal history that is separate from the other populations. This is more dramatically demonstrated in the MDS scatter plot, where the Kucong were eliminated as an outlier in order to produce a clearer differentiation between the other populations.

The possible explanation for this finding is that the Kucong have a more ancient, restricted maternal ancestry compared to the other eight populations. This is evident from the scant available demographic evidence which describes the Kucong as having only recently relocated into villages and towns from the Yunnan/Indochina border regions (Yang *et al.*, 1994).

Figure 5.5 Unrooted Neighbour-Joining Tree based on mean Tamura-Nei genetic distances calculated for mtDNA HV1 sequence data from the nine study populations

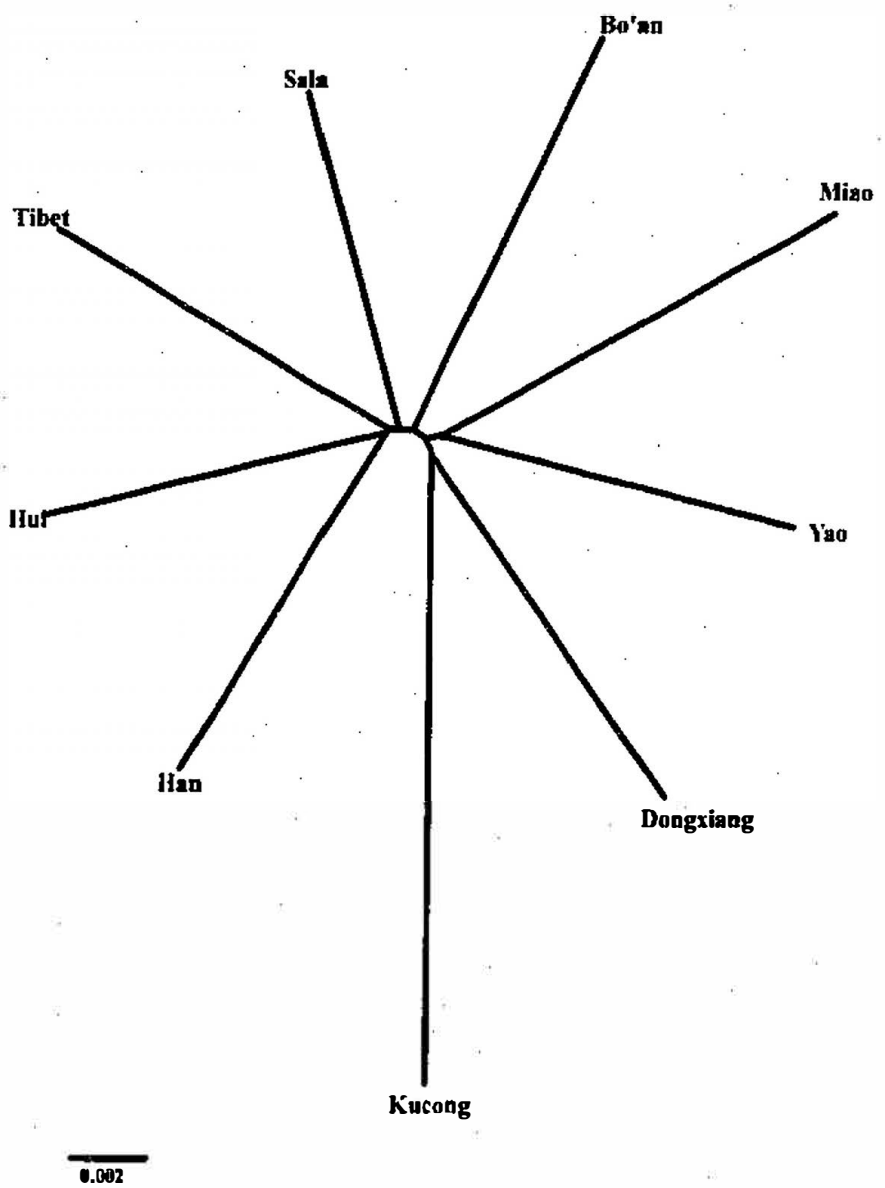
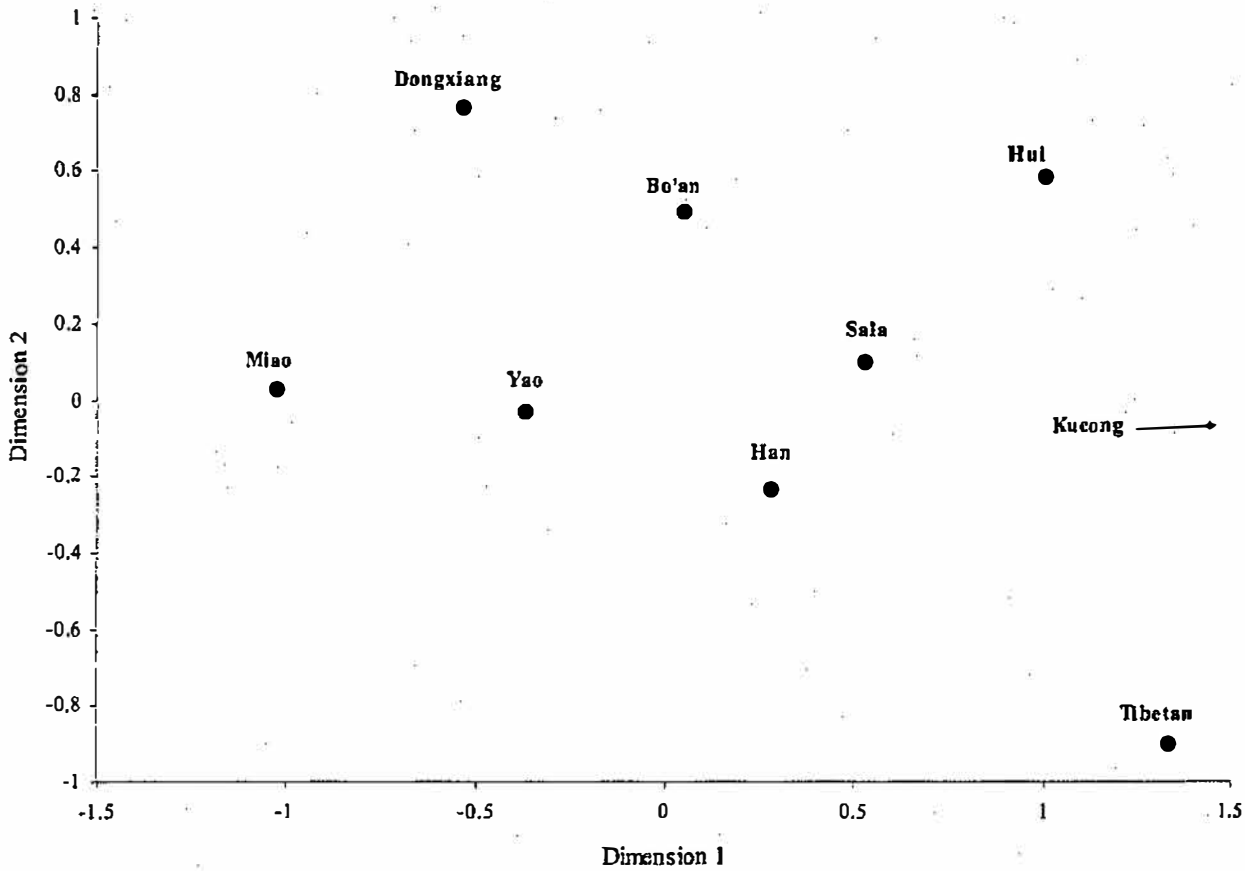


Figure 5.6 Multidimensional Scaling Analysis scatter plot based on genetic distances calculated mtDNA HV1 sequence data from the nine study populations



5.5 Summary

The conclusions that can be reached from this broad survey are that it is possible to discern different aspects of the genetic history of the study populations by simultaneously analysing autosomal, Y-chromosome and mtDNA data. For example, the finding of diverse paternal genetic structure is in line with historical accounts, especially for the 'Recent' populations, which have multiple ancestries from Central, West and Northern Asia. In contrast, with the exception of the Kucong, the maternal gene pools of the populations demonstrate high levels of homogeneity and thus the possibility of a largely shared maternal gene pool.

By utilising autosomal STR markers, it can further be demonstrated that the Chinese Muslim populations have a restricted gene pool, in keeping with accounts of long term endogamy within these populations. Conversely, the Han of Liaoning show similar results, but this is probably due to unrecognised population sub-structure (Black *et al.*, 2001). This demonstrates the influence of the genetic history of these populations on present-day genetic diversity, and that besides ancestral origins, events since their formation also are important in determining inter-population differences.

The data also indicate the possibility that the Kucong can be differentiated from the other populations on the basis of a more ancient ancestry. However, only broad assertions can be made from this summary statistic-based approach, especially in terms of a determination of ancestry and origins. To further explore the hypothesis of more ancient lineages in the Kucong, more detailed exploration of mtDNA and Y-chromosome ancestry is needed.

CHAPTER 6

GENEALOGICAL HAPLOGROUP ANALYSIS

6.1 Introduction

The analysis of population structure via summary statistical approaches revealed relationships between the study populations that could largely be ascribed to differences in internal population structure, rather than providing any direct inference of ancestry. For the inference of ancestry, one needs to utilise phylogeographical approaches developed from mtDNA and Y-chromosome biallelic markers. In this section the results of inference drawn from or based on haplogroups are presented, revealing an intriguing picture of diverse ancestral origins.

6.2 Mitochondrial haplogrouping

The apparent homogeneity of mtDNA diversity amongst the study populations revealed in Chapter 5 was further explored by assigning samples to haplogroups as defined in section 4.11.3. The division of haplogroups into two broad 'limbs' follows Kivisild *et al.* (2002), who defined major groupings as 'limbs' and sub-groupings as 'twigs' of the haplogroup 'tree'. All haplogroups outside Africa fall into either the M or N 'limbs', with a further branching from N at the R 'limb'. When mutations are discussed they are considered by comparison with the Cambridge Reference Sequence (CRS), which itself is northern European in origin and belongs to the HV haplogroup (Andrews *et al.*, 1999).

From the results displayed in Table 6.1, most sequences could be defined as belonging to haplogroups of East Asian origin. Haplogroups D, C, E and Z, all sub-groups of the M 'limb', are believed to be of East Asian origin (Kivisild *et al.*, 2002; Yao *et al.*, 2002a), and this is true for Group A on the N 'limb'. Haplogroups B and F are seen most commonly in populations in southern China and South East Asia and they are sub-groups of the R* 'limb', which in turn branches from the 'N' limb (Su *et al.*, 1999; Su *et al.*, 2000; Wen *et al.*, 2004b). Some samples could only be grouped into broad M* and R* haplogroups from HV1 motifs, and coding region polymorphisms would be needed for further haplogroup definition. This is important as recent studies have revealed many divergent M* and R* subgroups (such as M8 and R9) in various East Asian populations (Kivisild *et al.*, 2002).

It is possible to discern subtle differences in maternal ancestries among the study populations. Firstly, haplogroup D was the dominant group in the Han, Hui,

Table 6.1 Table of mtDNA haplogroup frequencies in the nine study populations

Population	No. of samples	M 'Limb'					N 'Limb'				
		D	M*	C	Z	E	A	W	B	F	R*
				'Recent' populations							
Han	11	0.500	0.167	0.083	-	-	-	-	0.167	-	0.083
Kucong	11	-	0.083	0.167	-		0.167		0.250	0.083	0.250
Miao	9	0.111	0.111	-	-	-	-	-	0.222	0.333	0.222
Tibetan	15	0.533	0.200	-	-	-	0.133	-	0.067	-	0.067
Yao	10	0.273	0.091	-	-	-	-	-	0.091	0.182	0.273
				'Ancient' populations							
Bo'an	14	0.091	-	0.273	-	-	0.455	-	-	0.182	-
Dongxiang	9	0.111	0.111	-	0.222	-	-	0.111	0.222	0.111	0.111
Hui	13	0.385	0.231	-	0.231	0.077	0.077	-	-	-	-
Sala	11	0.455	0.182	-	-	-	-	-	0.273		0.091

M* - Samples designated as M but no further definition could be made

R*- Samples designated by the absence of a 16223 mutation from CRS, and inability to further group

Sala and Tibetans, and it was present at low frequencies in all other populations, with the notable exception of the Kucong where it was absent. The Han, Hui, Sala and the Tibetans all have very different histories, yet the high frequency of D haplogroups in these populations suggests some common maternal origins. In fact, similarities in the types and frequency of haplogroups between the Han and Hui of Liaoning confirm the hypothesis of a shared maternal ancestry (Black *et al.*, 2001).

While no D haplotypes were inferred from the Kucong, it was found that the majority of Kucong samples were defined on the R limb (B, F, R*). Of those on the R 'limb', almost half belonged to haplogroup B while an equivalent proportion could not be further defined. Haplogroup R* represents all samples that had no 16223 mutation from CRS and could not be classified as either B or F. The R* samples of the Yao, Tibetans and Kucong did exhibit mutations at 16298 and 16362. According to Kivisild *et al.*, (2002) this is a motif for R9a. Such sub-R groups have been previously reported in Yunnan ethnic populations (Qian *et al.*, 2000, Yao *et al.*, 2002b and Yang *et al*, 2005). The R* present in the Dongxiang could be tentatively classified in the JT grouping which is largely Middle Eastern in origin and occurs at significant frequencies in some Central Asian populations (Zerjal *et al.*, 2002). Haplogroup W has a similar Middle Eastern origin, and again was found in the Dongxiang. It is most probable that these maternal lineages represent relatively recent migrations into the population. However, further anthropological investigation would be needed to confirm this hypothesis. One Han sample could also be tentatively be classified in the JT grouping and finding of this group in the Han is not unprecedented (Yao *et al.*, 2002a).

Haplogroup B was present in all populations except the Bo'an and Hui. However, it was only present at significant frequencies in the Kucong, Miao, Sala and Dongxiang. Haplogroup F was most frequent in the Bo'an, Miao and Yao. Both haplogroups F and B have been previously resolved as most frequent in South East Asian populations and among Yunnan ethnic populations (Yao *et al.*, 2002b, Thangaraj *et al.*, 2005, Trejaut *et al.*, 2005 and Yang *et al.*, 2005). Their significant presence thus shows the strong influence of South East Asian maternal ancestries in the Miao, Kucong, and Yao.

6.2.1 Principal Components Analysis of mtDNA haplogroups

To assess the overall level of haplogroup diversity within the study populations, and thus examine the hypothesis of shared East Asian maternal ancestries, the haplogroup frequency data were subjected to a Principal Components Analysis (PCA). Examination of the PCA scatter plot shows that the Sala, Han, Tibetans, and to a lesser extent the Hui, clustered together, probably due to their high frequency of D haplotypes. These four populations have very different recorded histories and yet have similar maternal ancestries, demonstrating the effect of migration and admixture of maternal lineages throughout East Asia (Ding *et al.*, 2000; Yao *et al.*, 2002a; Yao and Zhang, 2002).

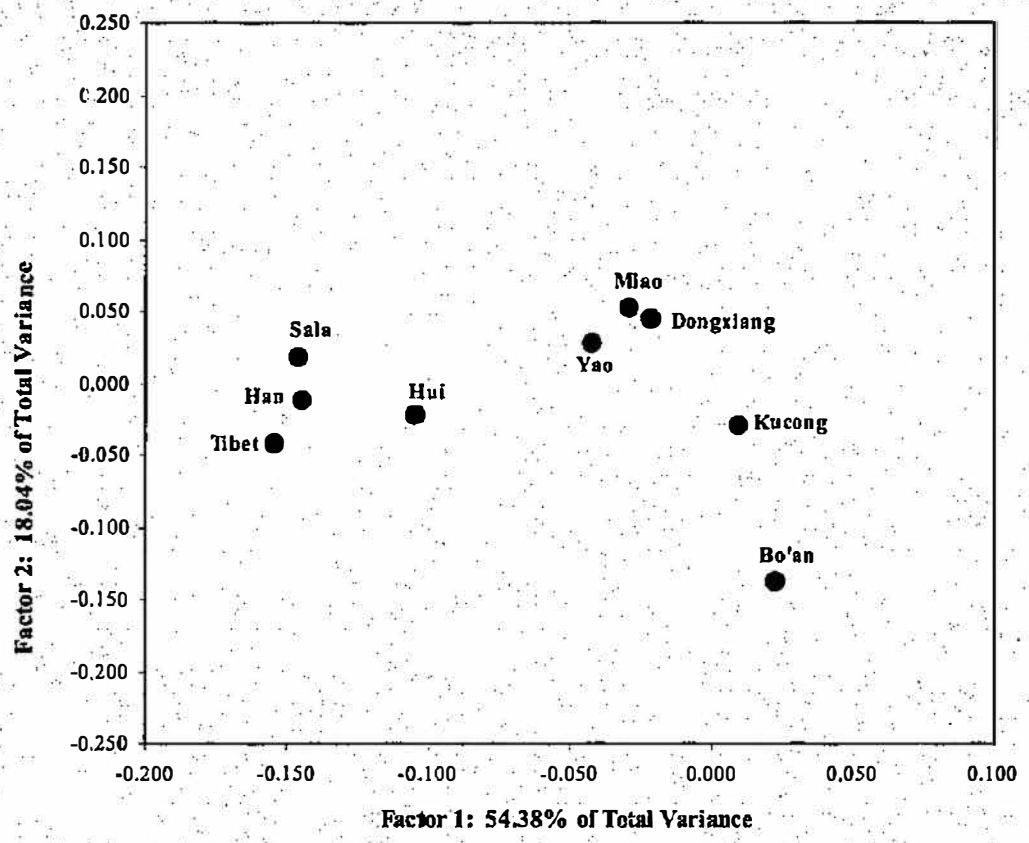
The position of the Miao, Dongxiang and Yao on the scatter plot is indicative of populations containing a significant frequency of both M limb haplogroups and R limb haplogroups. By comparison with previous studies one can conclude that these populations have at least two separate mtDNA ancestries; from northern East Asia (M limb) and southern East Asia (R limb) respectively (Yao and Zhang, 2002).

Inspection of Figure 6.1 shows that the Kucong are aligned between the Dongxiang and the Bo'an. Initially this seems counter-intuitive as from Table 6.1 it is clear that the Kucong mostly comprise R limb haplogroups, much like the Dongxiang, Miao and Yao, and thus should be more closely clustered with them. However, by comparison to the Bo'an, it is clear that the pattern of clustering results from the presence of haplogroups A and C in the Kucong, haplogroups which predominate in northern East Asian populations. This leads to two possible conclusions. Firstly, that the presence of these haplogroups is due to recent admixture, and thus the hypothesis that the Kucong are an ancient isolated population is at least partially inaccurate, or that these haplotypes are representative of more ancient A and C lineages than are present in the Bo'an. Clearly, a resolution of Y-chromosome lineages is needed to help clarify the ancestry of the Kucong.

Another intriguing result from this analysis is the wide separation of the three populations which are co-resident in Gansu province, the Dongxiang, Bo'an and Sala. The isolation of the Bo'an from the other two populations is due to their high frequency of C and A haplogroups, which are most frequent in northern East Asian populations. By their close clustering with the Miao, the Dongxiang show more of a

southern Chinese ancestry. Finally, the Sala are clustered close to the Han due to their predominant frequency of D haplogroups. In essence, one can conclude that the Bo'an and the Sala have largely separate northern East Asian maternal ancestries while the Dongxiang contain a more southern maternal ancestry. Given the small sample sizes, this interpretation must be treated with due caution and further sampling would be required before a definite conclusion could be drawn. Nonetheless, the mtDNA analyses are in keeping with historical sources which indicate that, despite living in close geographical proximity since their formation as recognised communities, there has been little admixture between these three Muslim populations (Du and Yip, 1993; Lipman, 1997).

Figure 6.1 PCA plot of the nine study populations based on mtDNA haplogroup frequencies



6.3 Y-chromosome haplogroups

The next stage of the ancestral analysis was to resolve paternal ancestries. Analysis of Y-chromosome markers for the 'Recent' populations and the Han is based on data from Wang *et al.* (2003) supplemented by additional new data resolved in this study. In the study of Wang *et al.* (2003), more extensive resolution of the UEP loci was achieved using MALDI-TOF analysis. However, in the current study only the main branching UEPs could be resolved because of the restricted amounts of DNA sample available for the Yao, Miao, Kucong and Tibetans. A summary of various haplogroup frequencies in each population is given in Table 6.2. It can be concluded that a significant difference in haplogroup diversity exists between the 'Ancient' and 'Recent' populations. Therefore the following discussion is divided into two parts according to these two groupings.

6.3.1 Haplotype diversity in the 'Ancient' populations

The majority of Han and Miao Y-chromosomes comprise haplogroup O and subgroup O3 haplogroup, indicative of shared paternal origins in East Asia. The haplogroup O family has previously been designated as East Asian in origin because of its high frequency in East Asian populations and virtual absence elsewhere. The high prevalence of O haplogroups is thought to result from a Neolithic population expansion following the establishment of rice and millet agriculture in the Upper and Middle Yellow River regions respectively about 10,000-8,000 YBP (Su *et al.*, 1999; Ding *et al.*, 2000; Deng *et al.*, 2004; Wen *et al.*, 2004b).

The high frequency of haplogroup O in the Miao indicates that they share male ancestry with the Han and in ancient Chinese texts it was indeed stated that the Miao had a common ancestral link with the Han (Diamond, 1995). The progenitors of both the Miao and another ethnic population, the Li, were said to have been defeated in a series of battles by the Huaxia, the legendary ancestors of the Han. As described in section 4.1.1.3, the Miao subsequently had a history of continuous migration southwards, until their arrival in the region now known as Yunnan Province (Du and Yip, 1993; Diamond, 1995; Yao *et al.*, 2002a). Thus the high frequency of O haplogroups in the Miao can probably best be attributed to a combination of shared paternal ancestry and subsequent admixture.

Table 6.2 Y-chromosome haplogroup frequencies in the nine study populations

Marker haplogroup[#]	No. of samples	YAP DE	M130 C	M89[^] F*(xK)	M175 O	M122 O3	M9 K*(xO,P)	M17 R1a1	M45 P*(xR)
				'Ancient' Populations					
Han	91	-	0.049	0.146	0.227	0.480	0.098	-	-
Kucong	21	0.190	-	0.524	-	0.286	-	-	-
Miao	32	-	0.053	0.132	0.282	0.533		-	-
Tibet	20	0.550	-	0.100	0.250	-	0.100	-	-
Yao	14	0.571	-	0.071	0.071	0.286	-	-	-
				'Recent' Populations					
Bo'an	40	0.056	0.028	0.167	0.056	0.250	0.111	0.250	0.056
Dongxiang	41	0.065	-	0.152	0.065	0.196	0.109	0.321	0.092
Hui	22	-	0.227	0.227	0.090	0.136	0.136	0.181	-
Sala	32	0.062	0.062	0.094	0.218	0.312	-	0.218	0.032

[#]Y Chromosome Consortium (http://ycc.biosci.arizona.edu)

[^]Markers M89 and M172 defined for Bo'an, Dongxiang, Hui, Sala and Han.
Only M89 defined for the Kucong, Miao, Tibet and Yao.

Haplogroup D represents another East Asian paternal ancestral group. Previously, this haplogroup has only been previously found at significant frequencies in Tibetan and Japanese populations (Torroni *et al.*, 1994; Qian *et al.*, 2000; Deng *et al.*, 2004; Wen *et al.*, 2004b), and at lower frequencies in some Yunnan ethnic groups (Deng *et al.*, 2004). In the current study the chosen designation was DE, based on the presence of a YAP insertion rather than any subsequent marker, which defines the split between D and E. From previous studies it appears certain that the existence of the polymorphism in the Tibetans, Yao and Kucong signifies haplogroup D. However the low frequencies of the insertion observed in the Bo'an, Dongxiang and Sala are more problematic. Haplogroup E, which also contains the YAP insertion, is present in populations resident in North Africa, the Mediterranean and at lower frequencies in Northern Europe and the Middle East. But, significantly, it has not been reported in East Asia (Semino *et al.*, 2004). Due to the migratory histories of the 'Recent' populations, it is possible that the existence of the YAP insertion could signify Haplogroup E. Unfortunately, insufficient samples were available to further define the possible presence of this haplogroup.

One important conclusion which can be reached is that the majority of Tibetan haplotypes were of haplogroup D as expected. The high frequency of this haplogroup in the Yao was, however, somewhat surprising and conflicted with the linguistic phylogeny which places the Miao and Yao in the Hmong-Mien language subgroup of the Sino-Tibetan language family (Grimes, 2003; Deng *et al.*, 2004). The Y-chromosome data indicated that while the Miao and Yao had a shared linguistic heritage, they did not have a shared paternal genetic heritage. On the basis of Y-chromosome haplogroups, it is inferred that both the Yao and Tibetans share an ancient male ancestry originating in North West China (Chu *et al.*, 1998; Su *et al.*, 1999; Deng *et al.*, 2004).

From Table 6.2 it is possible to conclude that the Kucong display a different, possibly more ancient, paternal ancestral heritage than each of the other study populations. This conclusion is based primarily on the fact that a majority of the Kucong samples resolved belonged to haplogroup F*(xK). Minor frequencies of this haplogroup were also found in the Yao, Miao and Tibetans. The existence of this haplogroup in the 'Recent' populations, as well as the Han, has been explained in terms of haplogroup J2, thus indicating West and Central Asian origin. The same

conclusion cannot, however, be drawn in respect to the Kucong. As previously described (section 4.2.1.2), the Kucong are composed of communities assembled from previously isolated hunter-gatherer groups from the Ailaoshan valley close to the China-Burma border. As such it would be unlikely that their dominant male paternal ancestry could be of recent West Asian origin.

In addition, the F(xK) haplogroup has been found at significant frequencies in Yunnan ethnic groups such as the Yi and Lahu, with whom the Kucong are unofficially listed as a subgroup (Xiaotong 1991, Yang *et al.*, 1994). Intriguingly, the presence of haplogroups O3 and D in the Kucong leads to the possibility that the current Kucong population may include tribes with different paternal histories.

In any event, it can be hypothesised that haplogroup F(xK) in the Yunnan ethnic populations is a remnant of ancient southern East Asian ancestry and does not result from the Neolithic expansion. If this hypothesis is correct, it could be further concluded that the Kucong themselves may represent a pre-Neolithic past.

6.3.2 Haplogroup diversity in the 'Recent' populations

In Table 6.2, the 'Recent' populations displayed a wider range of Y-chromosome haplogroups than the 'ancient' populations, encompassing a majority of known major haplogroups resolved from populations found throughout Eurasia. This wide diversity is in keeping with the historical evidence which reported that the male founders of these populations were merchants, soldiers and immigrants from West, North and Central Asia. This mixture of different ancestries is evident in their haplogroup data.

Interestingly, the numerically smallest of the 'Recent' populations, the Bo'an, displayed the most diverse distribution of haplogroups, comprising all haplogroups resolved in the study. To put this result into perspective, the Bo'an have a total population of 16,500 people and therefore most probably no more than ~8,000 males. The widespread pattern of Y-chromosome variation in this small population provides strong evidence that the Bo'an originally formed from diverse paternal founder populations sourced from individuals originating in West, Central and North Asia.

The demonstration of wide paternal diversity in all of the 'Recent' populations was reinforced by the fact that P*(xR) and R1a1 haplogroups were only

resolved in these populations. Previous studies have shown that these haplogroups originated in, and are most frequent among, populations in West, North and Central Asia (Karafet *et al.*, 2001; Wells *et al.*, 2001; Zerjal *et al.*, 2002, Wise *et al.*, 2005). Additionally, haplogroup R1a1 is believed to have originated in Central Asia 30,000 years ago and been dispersed within nomadic populations of the Eurasian steppe, aided by domestication of the horse around 7000-5000 YBP (Wells *et al.*, 2001). In fact, its direct ancestor R1* (M173) is the most common Western European Haplogroup, while M17 itself is virtually absent (Semino *et al.*, 2000). This is further proof of M17 more recent origins (Wells *et al.*, 2001). Therefore, the existence of P*(xR) and R1a1 haplotypes in the 'Recent' populations provided confirmatory evidence of more recent paternal ancestry stemming from populations from West, Central and possibly North Asia.

Further evidence of northern Asian ancestry is revealed by the resolution of haplogroup C, which is most prevelent in the Hui. This is an ancient haplogroup, believed to have arrived in the first migrations of modern humans into East Asia along a southern dispersal route (Underhill *et al.*, 2001). However, it is now only found in significant frequencies in Eurasia amongst Mongolians and Koreans (Wells *et al.*, 2001; Zerjal *et al.*, 2002). The high frequency of this haplogroup in the Hui thus signifies a strong northern Asian component of paternal origin, possibly Mongolian.

Finally, appreciable frequencies of F*(xK) haplogroup were present in the 'Recent' populations. Wang *et al.* (2003) had confirmed the M172 mutation via MALDI-TOF mass spectrometry, indicating their classification as haplogroup J2. Haplogroup J2 and its 'parent', haplogroup J, is found at highest frequencies in the Middle East, where it is thought to have arisen (Quintana-Murci *et al.*, 2001, Underhill *et al.*, 2001). It is also present at lower frequencies in North Africa and Europe (Semino *et al.*, 2004), in some Central Asian populations (Zerjal *et al.*, 2002), and in Indian populations (Kivisild *et al.*, 2003). It is, however, rare in East Asia and so its presence in the 'Recent' populations is most probably due to historical migration from West and Central Asia.

6.3.3 Principal Components Analysis of Y-chromosome haplogroups

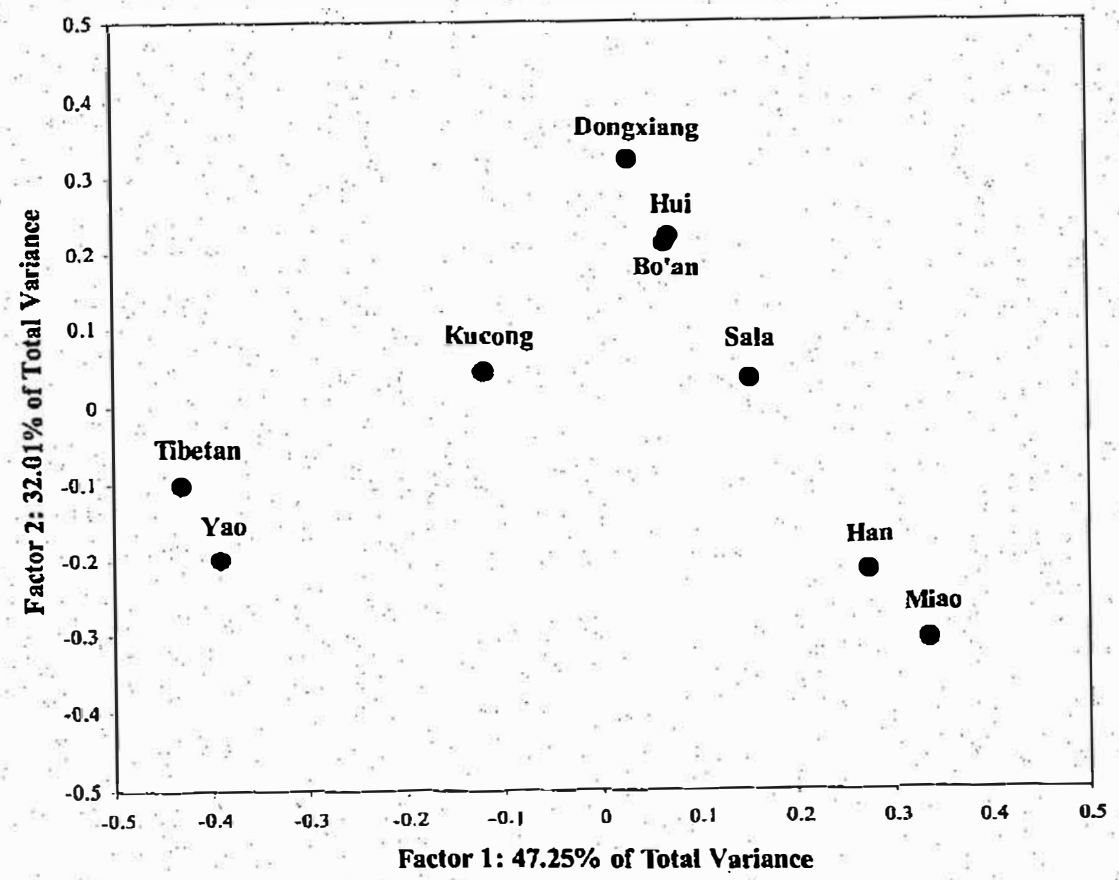
Further resolution of the different paternal ancestries of the study populations can clearly be discerned in a PCA scatterplot (Figure 6.2). The plot reveals a distinctive grouping in which the Bo'an, Dongxiang and Hui have equally diverse paternal ancestries, while the Tibetans and Yao have closely related ancestries, as do the Han and Miao. The plot is thus in keeping with the previous interpretation drawn directly from the data frequency table (Table 6.2). The isolated position of the Kucong is also in keeping with the previous interpretations of a distinct, ancient paternal ancestry when compared to the other eight study populations.

Perhaps the most interesting observation from Figure 6.2 is the apparent separation of the Sala from the other 'Recent' populations, due to the majority haplogroups O and O3 in the Sala samples. This finding initially seemed to contrast with the previous assumption that the Sala were largely of West, North and Central Asian paternal ancestry. However, a review of previous studies shows that these haplogroups have been found in low frequencies in a number of other Central Asian populations, such as the Dungan of Kyrgyzstan and various Uzbek tribes (Wells *et al.*, 2001). The high frequency of these haplogroups in the Sala could thus result from the effective return of the various O lineages back from Central Asia to East Asia, and be an artefact of the selective sweep effect of the East Asian Neolithic expansion. This tentative conclusion is strengthened by evidence indicating the occurrence of gene flow from East Asia to geographically adjacent Central Asian populations (Zerjal *et al.*, 2002), and the fact that historical sources consistently deny substantial conversions of local Han communities to Islam (Lipman, 1997; Gladney, 1998)

6.3.4 Network analysis

The Y-chromosome haplogroup diversity found among the study populations was further explored by examination of microsatellite haplotype diversity within each haplogroup. This was accomplished by median-joining network analysis. The approach to network construction followed the method recommended by Kayser *et al.*, (2001b), where a single-repeat mutation mechanism is assumed and each microsatellite marker is weighted based on the allele frequency of each locus within

Figure 6.2 PCA plot of the nine study populations based on Y-chromosome haplogroup frequencies



the haplogroups and their estimated mutation rates. The final weightings used were, DYS19, 1; DYS388, 0.1; DYS389I, 0.5; and DYS393, 1

All network inferences were further processed under the axiom of maximum parsimony available in the latest NETWORK software release. This process results in a 'cleaner' network, in which higher-dimension cubes and similar structures are avoided. These structures occur in Y-chromosome STR networks because of ambiguous branching, due to a lack of data and to recurrent mutations.

A convoluted pattern of recurrent mutation was found with the DYS388 STRs in haplogroup F(xK). By reference to previous studies, it was found that in this haplogroup, especially in sub-haplogroup J*, DYS388 displayed strong non-SMM mutational behaviour with recurrent multi-step mutations (Quintana-Murci *et al.*, 2001; Semino *et al.*, 2004). Hence, for haplogroup F(xK), DYS388 was not included in subsequent network analysis.

6.3.4.1 Maximum Parsimony Median Network Diagrams

Median Joining networks for haplogroups DE, F(xK),O3 and R1al are shown in Figures 6.3a-d. In these networks it was possible to cluster different haplotypes according to common DYS19 loci. While this may be an artefact of the weighting system used, it does permit a broad inference of the ancestral age of the populations to be drawn.

Haplogroup DE is dominated by a clustering of exclusively Tibetan and Yao haplotypes, all with the DYS19 15 repeat allele (Figure 6.3a). This close clustering within four mutations of each other is further evidence of their previously hypothesised shared ancestry, stemming from Northwest China. The restriction to just DYS19 15 could be construed as resulting from the extinction of many haplogroup D lineages consequent to the rapid spread of haplogroup O3. As a caveat, further sampling would be required in order to eliminate the possibility of small sample bias.

Interestingly, the lone Kucong haplotype contained a 13 repeat DYS19 allele, 2 repeats shorter than the Tibetan and Yao haplotypes. While the possibility of subsequent back mutation cannot be ruled out, the finding could indicate that the Kucong have a more ancient origin than the Tibet and Yao lineages. The Bo'an,

Dongxiang and Sala samples all displayed DYS19 14 alleles, yet they had widely different alleles for the other markers, thus indicating population specificity. However, as previously discussed, these haplotypes could be derived from haplogroup E and further investigation would be required to ascertain if this is the case.

Inference from haplogroup F(xK) also showed a pattern of clustering according to DYS19 repeat size (Figure 6.3b). As in haplogroup DE, the Kucong haplotypes displayed the smallest DYS19 repeat length of 13, however they also had haplotypes containing 14-repeat length DYS19 alleles. On the whole, this finding can be considered as further evidence of the previously posited ancient paternal history of the Kucong.

In fact, DYS19 14 was the most common DYS19 repeat in this haplogroup, and it was also present in the Han, Sala, Bo'an, Dongxiang, Hui and Tibetans. As has already been stated, these haplotypes in the Han, Bo'an, Dongxiang, Sala and Hui were actually resolved as belonging to sub-haplogroup J2. Therefore a direct comparison with the Tibetan and Kucong haplotypes cannot be made. By considering just the J2 haplotypes, some possible common paternal ancestry could be inferred, due to the sharing of DYS19 14 haplotypes between the Han, Dongxiang and Bo'an. However, because of the limited number of STRs surveyed, caution is needed in making this type of inference and the resolution of further markers may well separate these individuals.

The construction of a network for the most common haplogroup, haplogroup O3, provided another aspect of Chinese paternal history. The network revealed no obvious population-specific clustering. However, there was tight clustering between haplotypes separated by only one to three mutations (Figure 6.3c). This finding lends weight to the idea of O3 being spread quickly through many populations in China, and in Central and North Asia via the proposed Neolithic population expansion. Furthermore, the Kucong O3 haplotypes all contained the smallest DYS19 allele, 13 repeats, as was found in haplogroups DE and F(xK), which again can be interpreted as evidence of their more ancient paternal ancestry.

Finally, the R1a1 network revealed a wide spread of haplotypes (Figure 6.3d). The most striking feature was the fact that only one haplotype was shared

between the populations, i.e., between the Bo'an and Dongxiang, and resolution using further markers may in time also separate these two populations. Since only four markers were surveyed in this analysis, the lack of haplotype-sharing provides confirmatory evidence that the co-resident Bo'an, Dongxiang and Sala communities had little paternal admixture since their formation, i.e., they each had opted for an endogamous reproductive strategy.

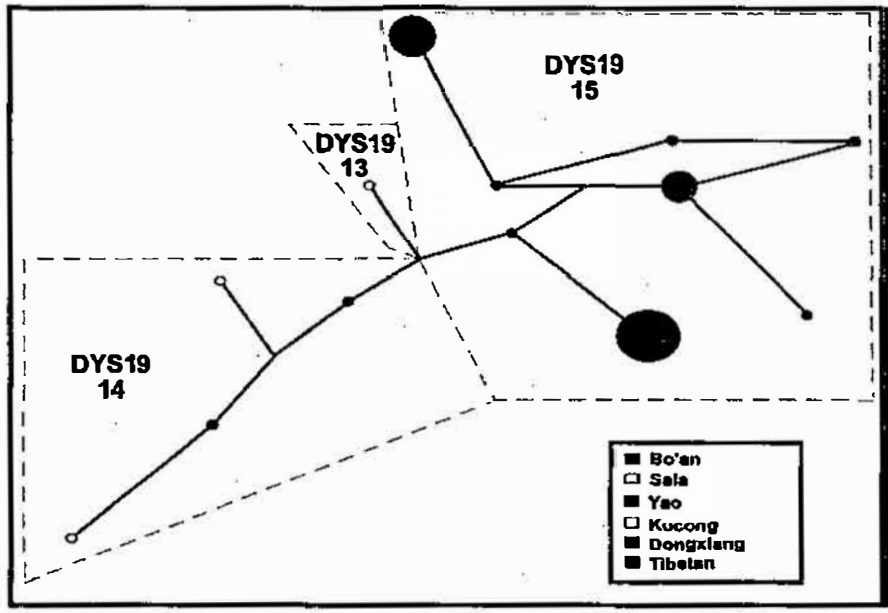
6.3.4.2 Rho statistics and TMRCA estimates

While inspection of the presence and size of clustering can be informative in assessing ancestral history, it is an empirical exercise which depends on individual interpretation. To establish a firm footing in inferring ancestral origins, a statistically significant measure is required. One such measure is the rho parameter (section 3.4.3).

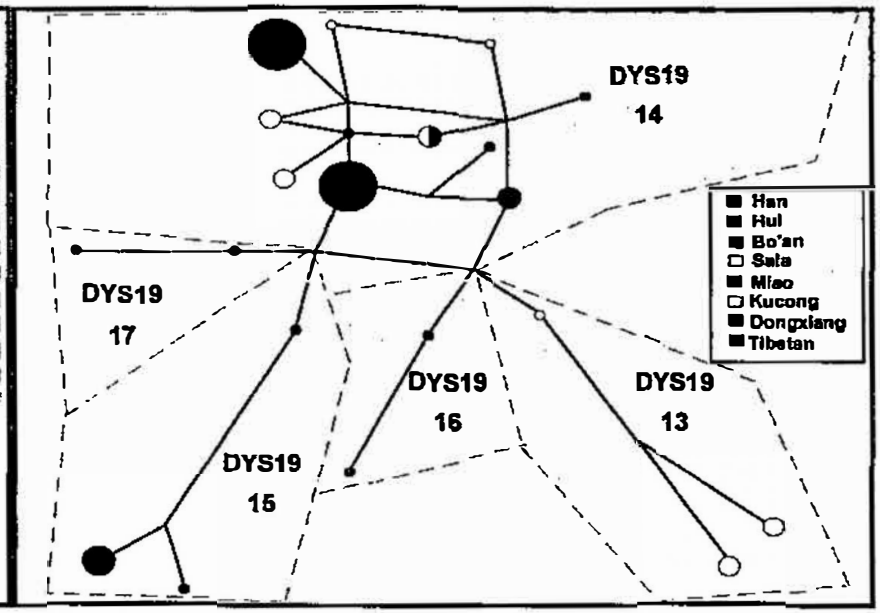
Calculation of the Rho statistic, and the associated standard deviation, was completed for all haplogroups except haplogroup P(R1). A value for TMRCA was calculated for each haplogroup using the effective mutation rate devised by Zhivotovsky *et al.* (2004). This rate takes into account both tri- and tetra-nucleotide repeats, as well as the possibility of multi-step mutations, which is especially important due to inclusion of the tri-nucleotide DYS388 in the analysis of most haplogroups.

Figure 6.3 Median Joining Networks of Y-chromosome STR haplotypes within selected haplogroups from the nine study populations

(a) Haplogroup DE



(b) Haplogroup F*(xK) including J2

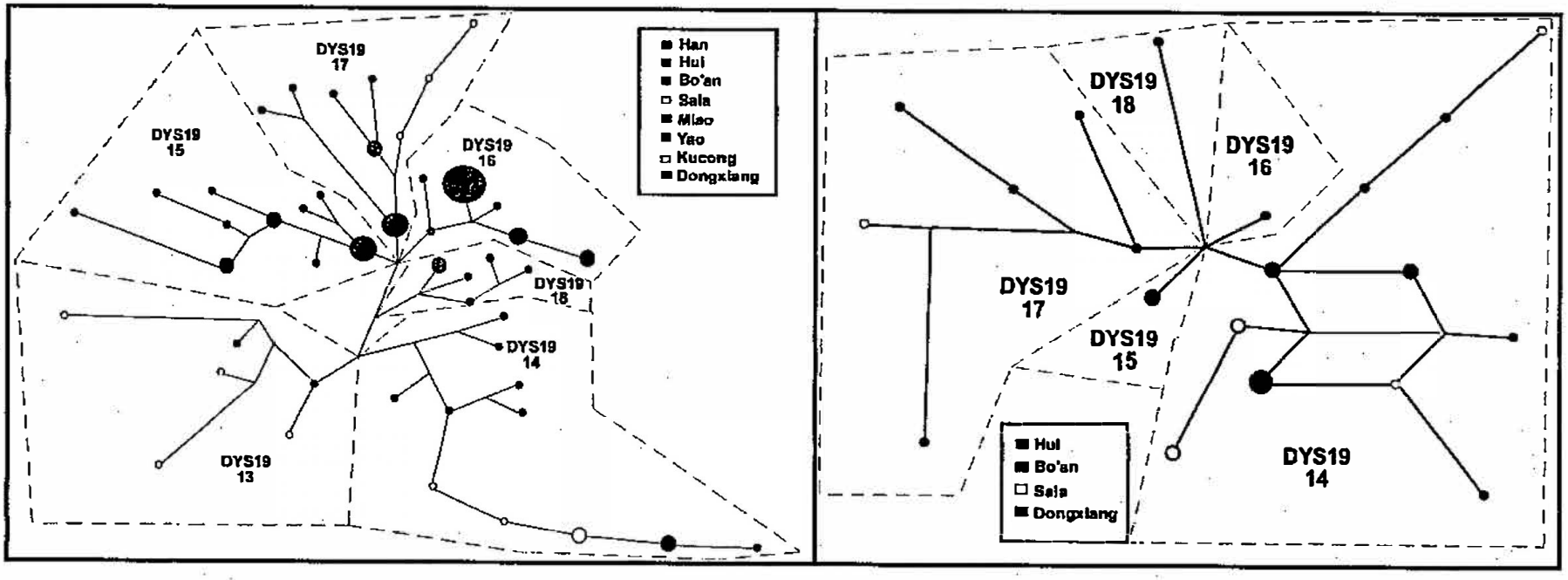


For clarity, STR mutation names and median vector nodes are not shown. Node size is proportional to haplotype frequency, with the smallest node equal to 1. Branch length is proportional to the number of repeat mutations. Grouping via green dotted lines indicate those haplotypes with a common DYS19 repeat.

Figure 6.3 (cont).

(c) Haplogroup O3

(d) Haplogroup R1a1



For clarity, STR mutation names and median vector nodes are not shown. Node size is proportional to haplotype frequency, with the smallest node equal to 1. Branch length is proportional to the number of repeat mutations. Grouping via green dotted lines indicate those haplotypes with a common DYS19 repeat.

Table 6.3 Rho statistics and estimated TMRCA of haplogroups resolved from the nine study populations

	ρ[1]	σ[1]	TMRCA (years)[2]		
			5% CI	Mean	95% CI
Haplogroup C	5.00	1.17	32,200	44,600	72,700
Haplogroup DE	3.45	1.10	22,200	30,800	50,100
Haplogroup F*(xK)	4.17	0.96	26,900	49,600	60,600
Haplogroup J2	3.20	0.97	20,600	38,000	46,500
Haplogroup K(xO,P)	3.50	0.84	22,500	31,250	50,900
Haplogroup O(xO3)	4.13	0.74	26,600	36,900	60,000
Haplogroup O3*	3.36	0.60	21,600	30,000	48,800
Haplogroup R1a1	3.50	0.88	22,550	31,250	50,900

[1] ρ statistic and standard deviation, σ, estimated from Median joining network. Age in generations calculated using the formula $\rho = t/\mu$ where μ is the mutation rate.

[2] Mutation rate of $6.9 \times 10^{-4} \pm 1.3 \times 10^{-4}$ per locus per generation (Zhivotovsky *et al.*, 2004). Final TMRCA values are rounded.

The estimated TMRCA values obtained are broadly in agreement with previous Y-chromosome based investigations which calculated that the first dispersal of modern humans out of Africa occurred 60,000 YBP, with a 95% CI of 40,000 to 140,000 years (Underhill *et al.*, 2001; Jobling *et al.*, 2004). Table 6.3 shows that haplogroups C and F(xK) are the oldest, recording TMRCAs of 49,600 and 44,600 (38,000 for J2) years respectively, in keeping with previous estimates (Underhill *et al.*, 2001). Interestingly, the K(xO,P) group had a lower estimated TMRCA than its 'progeny' haplogroups O and R1a1. This indicates that further resolution of markers for these samples is required, as these haplotypes may be classified in the progeny haplogroups L, M or N rather than in the ancestral K* haplogroup.

Haplogroup DE recorded a TMRCA of ~31,000 YBP, a figure well below the TMRCA of haplogroup C and which conflicts with the hypothesis that these haplogroups arrived more or less simultaneously with the first migrations into East Asia. The lower value for haplogroup DE is reflected in the limited diversity displayed in the DE haplotypes of the Tibetans and Yao. If only the Tibetan and Yao samples in the network are considered a ρ value of 2.41 is found, resulting in a

TMRCA of just 21,500 years (95% CI, 15,500-34,000). This could possibly be indicative of the hypothesised arrival of the K and O haplogroups into China 40,000-30,000 YBP, subsequently replacing most D and C haplotypes and extinguishing many ancient lineages (Underhill *et al.*, 2001).

However, all of the TMRCA estimates are very broad due to the small number of loci used. Small Y-chromosome haplotypes can be highly variable and result in large confidence intervals (Pritchard *et al.*, 1999). Therefore hundreds of loci may be needed for precise dating of ancient demographic events (Goldstein *et al.*, 1996; Zhivotovsky *et al.*, 2004). From these observations, time-based estimates made from data resolved in the present study must be treated with caution.

Nonetheless, the TMRCA values calculated do broadly concur with the hypothesised evolutionary history of the haplogroups, including selective sweeps of O haplogroups and the virtual extinction of more ancient lineages in East Asia. More importantly, they help to further confirm the hypothesis that the Kucong have a more ancient paternal ancestry due to the fact the majority of Kucong belong to the F*(xK) haplogroup, which is inferred as the most ancient (Table 6.3).

6.3.5 Combining inter- and intra-population structures

The next step in the analysis of haplogroup diversity was to investigate the nature of this diversity within each study population, rather than between the haplogroups. This investigation was restricted to haplogroups O*, O3, R1a1 and F*(xK), as there were only statistically sufficient STR haplotype data available for each population within these haplogroups. For the purpose of population comparison, the genetic distance $(\delta\mu)^2$ was calculated. As in section 5.3.4, MDS proved to be more informative than phylogenetic analysis, and so only MDS was performed on the distance matrices produced for the current analysis. Populations for which only a small number of haplotypes (<5) were present in a haplogroup were eliminated from the analysis to increase the level of statistical significance. The results of the analyses are shown in Figures 6.4 a-d.

It is apparent that there is no clustering in haplogroup F*(xK). For the Bo'an, Dongxiang, Hui and Sala this could be inferred as indicative of the effect of demographic processes that have occurred since the populations formed, such as

recent migration and endogamy. For the Kucong no such conclusion can be drawn because of the near-total lack of evidence of any such processes in this population, and the fact that, unlike the other study populations, further haplogroup definitions could not be made. Similarly, population-specific demographic processes also may influence haplogroup O which, as shown in Figure 7.4b, also shows little population clustering.

Haplogroup O3 exhibits a very different pattern to haplogroups F(xK) and D (Figure 6.4c). Firstly, it is evident that there is tight clustering of the Bo'an, Dongxiang, Han, Hui, Miao, Sala and Yao populations. This provides further evidence of the previously discussed selective sweep effect of the Neolithic population expansion in East Asia, and thus a common origin for these haplotypes in what otherwise are genetically diverse populations. Secondly, there is a distinct separation of the Kucong from the other populations. Previous inspection of the haplotypes belonging to haplogroup O3 showed that the Kucong samples all have the smaller DYS19 13 repeat allele, while the remaining eight populations exhibit a mixture of different DYS19 repeat sizes. Therefore, it could be concluded that the Kucong have either more limited and/or more ancient O3 haplotypes than the other populations. It could further be hypothesised that the Kucong harbour STR haplotypes that are remnants of ancient lineages going back to the Neolithic, while the haplotypes present in 'Recent' populations have been influenced by subsequent population expansion and admixture. The fact that O* and O3 are present in the 'Recent' populations can be construed as evidence of the spread of genes, along with the spread of agriculture, to populations outside the Yellow River Basin.

Finally, the MDS plot of genetic distances between populations within the R1a1 haplogroup is shown Figure 7.4d. This plot indicates a clear separation between the Bo'an, Dongxiang and the Sala along the first dimension. This could well be due to the presence of two different paternal ancestral histories, with the Dongxiang and Bo'an reputed to have Mongolian origins while the Sala are said to be Turkic (Du and Yip, 1993; Wang *et al.*, 2003). So the large genetic distance between these quite recently founded populations can be regarded as additional evidence of their proposed disparate ancestries and subsequent endogamy.

Figure 6.4 Multidimensional Scaling Analysis based on genetic distances calculated for Y-chromosome microsatellite data from the nine study populations

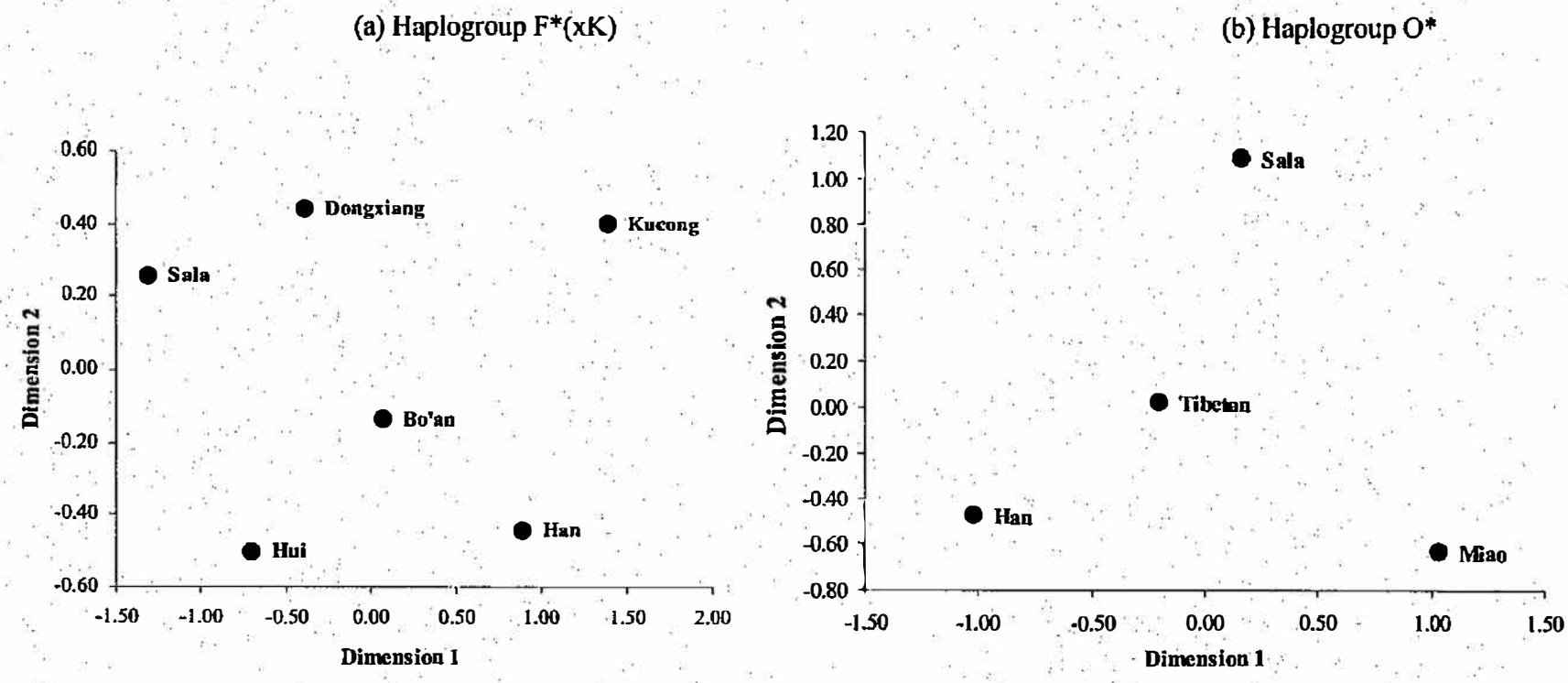
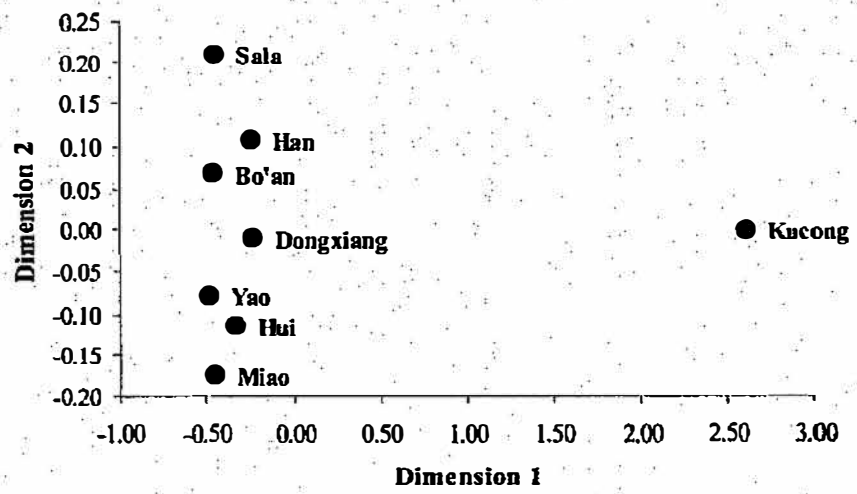
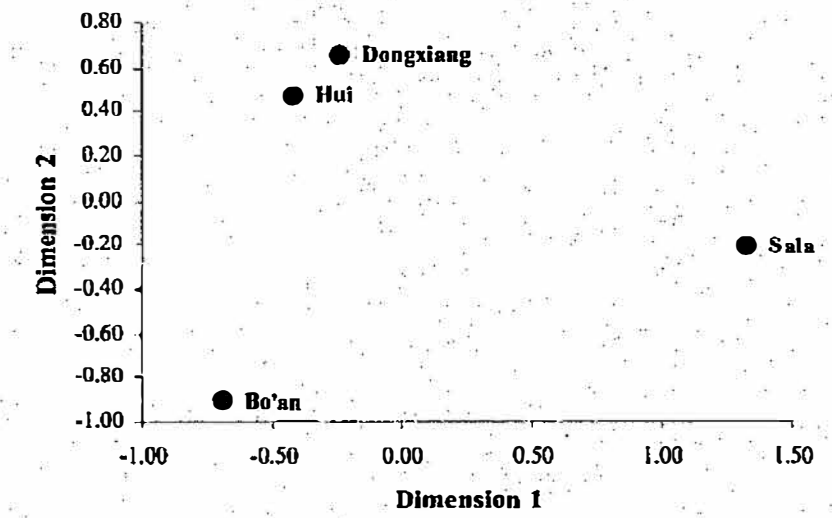


Figure 6.4 (cont.)

(c) Haplogroup O3



(d) Haplogroup R1a1



6.4 Summary

Resolution of both maternal and paternal haplogroups has demonstrated that, even from small samples, one can elucidate a picture of complex genetic structure in the present-day population of PR China. From a maternal perspective, it is possible to differentiate between northern and southern East Asian maternal ancestries in the study populations. Importantly, several findings of ancestral origin have been established that were not evident from the summary-statistic based mtDNA analysis, such as the large differentiation of the Bo'an, Dongxiang and Sala, and a possible northern East Asian ancestral component in the Kucong.

Y-chromosome haplogroups reveal ancestral origins largely in agreement with previous hypotheses, although the presence of a high frequency of haplogroup O3 in the Sala and lower frequencies in the other 'Recent' populations was not expected. This provides evidence for the strong effect imposed by the Neolithic population expansion on populations in East and Central Asia. The study also has resulted in the resolution of genetic evidence for wide ancestral origins of the 'Recent' populations, and possible pre-Neolithic origins for the Kucong.

The limited size of the available data set has resulted in some restrictions to the conclusions that can be drawn. For example, only broad haplogroup ages could be inferred, and these times have little bearing on the current populations in terms of their history and structure. To confirm inferred population histories, a more in-depth analysis is presented in Chapter 7 using likelihood-based stochastic coalescent modelling.

CHAPTER 7

LIKELIHOOD-BASED STOCHASTIC COALESCENT MODELLING

7.1 Introduction

To further investigate the genetic structure of the different populations, methods are required that allow an alternative approach to interpreting the complex interacting genetic processes that have occurred. The approaches now being developed are based on estimating and statistically testing fundamental population genetic parameters using likelihood and probabilistic modelling, stemming from the development of coalescent theory (section 3.5). Only recently with the development of sufficient computational power has it become feasible to construct and explore such models.

Three computationally intensive methods were utilised in the investigation of the genetic ancestry and structure of the study populations. Firstly, a broad assessment of autosomal, mtDNA and Y-chromosome diversity in each population was conducted by inference of effective population size using the MIGRATE-N algorithm. This algorithm is a maximum-likelihood estimation-based technique for exploring population migration models. The basic likelihood model from which it is constructed was briefly discussed in section 3.5.3. This was followed by an in-depth investigation of the autosomal genetic structure of the study populations via the STRUCTURE method, a set of Bayesian genotypic population assignment algorithms (section 3.5.5). Finally, as previous results had shown that the major differences in the ancestral diversity of the study populations were paternal in origin, an in-depth investigation of paternal ancestry was undertaken using the BATWING algorithm, to infer multiple parameters, such as ancestral population size, growth rate and TMRCA.

The three techniques are by no means the only methods available, but they are representative of the more popular likelihood-modelling based techniques that have been described in recent published research (Wilson *et al.*, 2001; Chaix *et al.*, 2004; Knowles, 2004).

7.2 Maximum-Likelihood Estimation of Genetic Diversity

MIGRATE-N is a maximum likelihood approach which infers effective population size, Θ, and migration rate, M, using the basic likelihood model shown in sections 3.2.3 and equation 3.24. More specifically, this model estimates the set of parameters, P, for n populations:

$$P = \begin{pmatrix} \Theta_1 & M_{21} & M_{31} & \dots & M_{n1} \\ M_{12} & \Theta_2 & M_{32} & \dots & M_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ M_{1n} & M_{2n} & \dots & M_{n-1,n} & \Theta_n \end{pmatrix}$$

Equation 7.1 Parameter Matrix for the MIGRATE-N Model

where $\Theta_i = 2hN_i\mu$ and $M_{ji} = m_{ji}/\mu$. The term h is the number of heritable copies of a locus per individual (i.e., 2 for diploid loci and 0.5 for haploid loci), N_i is the effective population size of population i, μ is the per locus mutation rate and m_{ij} is the immigration rate from population j to population i.

As was shown in equation 3.27, the calculation of likelihood is achieved by estimating a ratio based on the probability of a genealogy given the population parameters $P(G|P)$. For MIGRATE-N, this probability is expressed using an extension of Kingman's coalescent which enables the recording of coalescences in different subpopulations and migration events that switch lineages from one population to another over all time intervals, T (Beerli and Felsenstein, 1999). Specifically it is of the form:

$$P(G \mid P) = \prod_{j=1}^{T} \left[\exp\left(-u_j \left[\sum_{i=1}^{s} \frac{k_{ji}(k_{ji}-1)}{\Theta_i} \right] + k_{ji} \sum_{z \neq i}^{s} M_{zi} \right) \cdot \left(\delta_j M_{w_j v_j} + (1-\delta_j) \frac{2}{\Theta_v} \right) \right]$$

Equation 7.2 Likelihood of the Genealogy given the Parameters of the MIGRATE-N Model

where u_j is the length of time interval j scaled by generations, the mutation rate, s is the number of subpopulations, and k_{jl} is the number of lineages in subpopulation I during time interval j. The indicator variable δ_j is 1 when the event at the bottom of interval j is a migration event and otherwise is zero. The parameter $M_{w_j v_j}$ represents migrations from subpopulation w_j to v_j.

While developed as a method for the inference of migration between populations, in the present study MIGRATE-N was primarily used to infer similarities between populations, on the assumption that these populations have not directly exchanged migrants because of their different historical, cultural and geographical situations. However, MIGRATE-N does have the advantage of being able to handle a wide range of data, including DNA sequences and STRs.

For all analyses, it was found that a Markov Coupled Markov Chain Monte Carlo method (MCMCMC or MC^3) was required rather than the default MCMC method, as for some simulations the phenomenon termed 'the fatal attraction to zero' occurred. This arises as the chains become 'stuck' near zero and are unable to adequately explore the sample space (Beerli and Felsenstein, 1999). This problem was especially prevalent for the mtDNA HV1 sequence data, probably because they exhibit more complicated mutational behaviour than the STRs.

The MC^3 method improves mixing, and therefore the exploration of sample space, by running multiple chains simultaneously comprising a 'cold' chain, and a number of incrementally 'heated' chains. Each chain has its own fully specified state and is independent of other chains. At each generation, proposals are made to swap the position of two random chains and at the end of the simulation the value of the 'cold' chain is recorded. As the method uses multiple chains and the simulation of each additional chain doubles the run time, it requires equivalently longer runs. For example, the average time needed for one MC^3 migrate run using mtDNA sequence data was approximately five days before post-output statistical analysis. In addition, multiple runs were needed to adequately account for random variation.

After extensive experimentation, optimum settings were determined for autosomal STR, Y-chromosome STR and mtDNA sequence data. For autosomal and Y-chromosome data, the search strategy, which was based on a combination of

software documentation, previous studies and experimentation, was one of 10 short chains of 10,000 steps followed by two long chains of 100,000 steps. Of course being a MC^3 run, each chain was shadowed by alternate chains, in this case three chains at 'temperatures' of 1, 1.5 and 3 respectively.

An advantage of the MIGRATE-N system is the ability to assess alternative mutation models. For the present study it was found that the ladder mutation model was most adequate for the STR data, rather than the strict SMM model. The ladder mutation model is an improvement on SMM as it allows for occasional multi-repeat mutations, thus enabling the researcher to take into account variations in the mutational behaviour of STR loci. This is especially pertinent given the non-SMM mutational behaviour of DYS388 in some circumstances (section 6.3.4). In addition, for the STR data the initial genealogy was defined by the default UPGMA method and initial values of the Θ and M parameters were estimated by the default F_{ST} method.

The modelling was more complicated for the mtDNA HV1 data. This region of the mitochondrial genome has unusual mutational properties, with individual nucleotides having differing mutational rates and hence the ability to broadly define haplogroups. A methodology similar to that developed by Chaix *et al.* (2004) was adopted. In this scheme, each nucleotide (A, C, G, and T) is given different relative substitution rates of 0.002, 0.090, 0.549 and 2.465 respectively, and a transition/transversion ratio equal to 25. As suggested by Beerli and Felsenstein (2001), the initial genealogy was constructed with the PHYLIP software (version 3.5) using the Kimura 2-parameter distance and assuming a transition/transversion ratio of 25. The starting values of the Θ and M parameters were estimated according to Watterson's method (equation 3.3), and the exponential growth rate starting value was set at 1.0. The search strategy adopted was a MC^3 method, where it was found that an extended run involving 10 short chains of 50,000 steps and three long chains of 500,000 steps were required for adequate convergence. In fact, it was initially difficult to find a MC^3 search strategy where stable convergence was attained. As will be discussed later, this became a critical point in the appropriateness of MIGRATE-N in the interpretation of mtDNA diversity among the study populations.

The parameter Θ was of primary interest in this analysis as the interpretation of M is problematic. Theoretically, M is an inference of direct migration between populations. However, it cannot be assumed that direct migration has occurred between any the populations as was revealed in historical evidence (section 4.1) and genetic evidence (Chapters 5 and 6). It has been stated that M values might be expressed as genetic distances which take into account variances and co-variances of allele frequencies (Beerli, 2004). However, exhaustive simulation tests by Abdo *et al.* (2004) showed that the estimation of M by MIGRATE-N can be unreliable in many situations. Thus for the present study a cautious approach was taken where M values were considered. The results of the estimation of Θ and M are shown in Tables 7.1-7.3.

The autosomal Θ values shown in Table 7.1 for the 'Recent populations' were significantly larger than for the 'Ancient' populations (Wilcoxon rank sum test, $p = 0.016$). Taking into account the fact that size difference is used in the mutational model, this suggests a wider range of alleles is present in the 'Recent' populations, a finding consistent with their founders being from diverse ancestries. A survey of M values (Table 7.1) showed a range between 0.456 (Dongxiang to Yao) to 73 (Hui to Sala). The distribution of M values does have an interesting pattern, with significantly larger values recorded for M values representing 'migration' between the 'Recent' populations than those measured between the 'Recent' and 'Ancient' populations ($p = 0.002$). This suggests that the genotypic distributions of the 'Recent' populations' can most convincingly be explained in terms of their similar population histories and with diverse founder ancestries, than the more restricted ancestry of the 'Ancient' populations.

Unlike the autosomal analyses, the Θ values for the mtDNA analyses (Table 7.2) showed no significant difference between the 'Recent' and 'Ancient' populations, with a p value of 0.412 from a Wilcoxon rank sum test. On separate inspection of the populations, the Miao and Yao displayed much larger Θ values than the other populations (0.059 and 0.05 respectively), which is possibly indicative of their diverse mtDNA ancestry from both northern and southern East Asia as demonstrated in Chapter 6. The simulation process for the mtDNA data was,

however, unstable and experimentation with a range of MC^3 conditions was needed before a process with chains reaching a reliable convergence could be obtained.

The wide range of values for *M* (Table 7.2), is a good example of this instability, with estimates ranging from 0.2 (Hui to Miao) to a value of 1298 (Yao to Tibet) and a median value of 152. The fact that many of the values were >100 could be construed as evidence of mtDNA homogeneity among the otherwise diverse study populations. Alongside the previous summary statistic results (section 5.2) this provides further evidence to support the hypothesised widespread admixture of Han females within different Chinese ethnic populations.

Nevertheless, the combination of unstable simulations, the known historical origins and associations of these populations, and the complicated nature of the mutational behaviour of mtDNA HV1 sequence, indicates that MIGRATE-N results cannot be regarded as conclusive and mtDNA coding region polymorphisms may be better suited for stochastic likelihood modelling. Either a model that can account for this level of complexity is needed or resolution of the mtDNA coding region polymorphisms is required before the MIGRATE-N approach can be utilized for mtDNA diversity studies.

With their more straightforward mutational behaviour, the investigation of Y-chromosome STR variability was more reliable (Table 7.3). The Θ values for virtually all study populations were similar (0.240 – 0.329) with the Sala as the only outlier (0.416), and the difference between the 'Recent' and 'Ancient' populations not significant ($p = 0.413$). The restricted range of calculated diversity values was similar to the gene diversity findings shown in Table 5.4, indicating similar Y-STR distributions. Furthermore, overall lower *M* values were found for the Y-chromosome data, compared to the mtDNA analysis, with values ranging from 2.199 (Sala to Han) to 6.78 (Han to Dongxiang) with a median value of 4.02. As found in the summary statistic analysis (Table 5.4) and the diverse paternal ancestral origins uncovered in Chapter 6, this appears to be in keeping with the hypothesis of a relative lack of male inter-population admixture compared to the maternal gene pools of the study populations.

It has been suggested that *M* values could demonstrate the presence of population differentiation in a manner similar to *F*-statistics (Beerli, 2004). Thus it

could be postulated that since the Y-chromosome estimates are lower than those for mtDNA, they are in keeping with the previous findings of relatively homogenous maternal but diverse paternal origins among the study populations.

More detailed modelling is, however, required before a definite conclusion can be reached. The MIGRATE-N model estimated only two parameters using a maximum likelihood framework, which it could be argued limits the amount of information resolved from the raw data. The MIGRATE-N approach proved both mathematically successful and historically meaningful for Chaix *et al.* (2004) in estimating the genetic diversity and age of three Bulgarian Vlax Roma communities. But the structure of these communities was known at the level of family pedigrees and thus presented data with a relatively straightforward genealogical model. As a result, Chaix *et al.* (2004) could proceed with the calculation of less complicated likelihoods and so were able to allow direct historical interpretation of the M parameter. In the current study, data from the diverse 'randomly' sampled populations were analysed in the absence of direct genealogical information. The resultant uncertainty resulted in likelihoods that were significantly more complex, with a more abstract interpretation of the M parameter required.

For pedigree- and kinship-based studies with moderate amounts of data, the MIGRATE-N method is a statistically powerful approach to the inference of population genetic history. For a more generalized study, a more flexible method is needed that would allow different population genetic parameters and models to reflect more complex genetic interactions and histories. In any case, together with previously derived data, the MIGRATE-N model indicated that the major genetic differences between the study populations mostly resided in their autosomal and Y-chromosome structures, which formed the bases of the following analyses.

Table 7.1 Matrix of Maximum Likelihood Estimates of Θ and M parameters for autosomal microsatellite data from the nine study populations

			M								
			Donating Population								
			'Ancient' populations					*'Recent' populations*			
		Θi	Han	Kucong	Miao	Tibet	Yao	Bo'an	Dongxiang	Hui	Sala
Receiving Population	Han	3.884	-------	5.447	7.591	5.341	7.064	8.841	8.156	7.943	8.295
	Kucong	2.707	2.652	-------	4.088	1.849	1.458	3.451	2.532	5.883	3.988
	Miao	2.452	3.881	3.382	-------	24.663	30.422	2.185	2.605	11.009	3.079
	Tibet	3.662	2.787	5.402	44.182	-------	2.686	1.479	2.683	8.006	4.769
	Yao	1.612	2.034	1.736	18.719	2.296	-------	2.155	0.466	3.569	2.284
	Bo'an	10.066	12.361	7.429	9.756	12.439	13.895	-------	16.753	41.949	14.715
	Dongxiang	6.945	10.823	9.678	9.003	8.578	8.168	8.454	-------	17.673	11.660
	Hui	14.416	29.861	26.334	39.348	16.072	18.268	38.225	30.016	-------	55.683
	Sala	8.927	21.101	29.236	21.936	14.201	11.708	21.363	21.777	73.683	-------

Table 7.2 Matrix of Maximum Likelihood Estimates of Θ and M parameters for mtDNA sequence data from the nine study populations

			M								
			Donating Population								
			'Ancient' populations					*'Recent' populations*			
		Θi	Han	Kucong	Miao	Tibet	Yao	Bo'an	Dongxiang	Hui	Sala
Receiving Population	Han	0.018	------	316.5	364.8	202.4	485.4	159.3	234.2	0.6	0.4
	Kucong	0.014	326.6	------	69.5	43	19.3	134.4	0.2	63.1	223
	Miao	0.059	38.5	155.3	------	163.7	230.7	0.7	36.8	0.2	150.4
	Tibet	0.023	434.4	154.8	130.7	------	1298.8	71.1	467.7	1207.1	734.3
	Yao	0.05	58.5	182	18.4	128	------	83.1	233.1	379.6	488.9
	Bo'an	0.021	631.1	0.6	288.8	1167.7	182.9	------	669.9	435.5	389.2
	Dongxiang	0.038	0.8	722.6	245.4	116.2	298.6	538.4	------	406.2	1.4
	Hui	0.012	136.4	1036.3	0.5	703.7	0.4	570	109	------	140.7
	Sala	0.016	52.6	12.7	29	0.3	35.6	8.9	9.7	0.2	------

Table 7.3 Matrix of Maximum Likelihood Estimates of Θ and M parameters for Y-chromosome microsatellite data from the nine study populations

			M								
			Donating Population								
			'Ancient' populations					*'Recent' populations*			
		Θi	Han	Kucong	Miao	Tibet	Yao	Bo'an	Dongxiang	Hui	Sala
Receiving Population	Han	0.269	-------	2.248	4.045	3.455	3.436	2.837	2.813	2.254	2.199
	Kucong	0.224	4.665	-------	5.168	2.622	5.199	4.289	4.232	4.317	2.970
	Miao	0.256	2.473	3.644	-------	4.762	4.186	4.611	3.458	5.000	2.773
	Tibet	0.242	3.379	3.712	6.243	-------	5.010	3.749	5.006	3.431	5.805
	Yao	0.261	4.403	2.763	3.340	5.507	-------	3.365	5.121	4.661	3.499
	Bo'an	0.257	4.758	5.497	4.659	4.561	6.325	-------	4.008	4.878	3.540
	Dongxiang	0.329	6.785	4.492	4.756	3.204	2.862	4.207	-------	2.718	3.703
	Hui	0.240	2.723	3.694	4.069	3.454	3.378	5.177	3.580	-------	3.387
	Sala	0.416	2.257	2.644	4.951	5.161	3.685	6.659	4.196	6.513	-------

7.3 Autosomal genotypic structure

A significant aspect of this study is the focus on internal population diversity using autosomal markers. The different autosomal genetic structures of the study populations were previously apparent in the summary-statistic analyses, which demonstrated high levels of homozygosity in concert with a large range of alleles in the Hui and Han, and to a lesser extent in the Bo'an, Dongxiang and Sala. To further explore the extent and nature of this genetic diversity, the Bayesian STRUCTURE algorithm was utilised. As detailed in section 3.5.5, this is a Bayesian population assignment algorithm developed by Pritchard *et al.* (2001) which has been extended to include correlated allele frequencies and linkage (Falush *et al.*, 2003).

The algorithm constructs genetic clusters from a collection of individual multi-locus genotypes, estimating for each individual the fractions of the genome that belong to each cluster. The algorithm was comprehensively utilized in a recent investigation of global genetic diversity (Rosenberg *et al.*, 2002). The present analysis is by no means as exhaustive and it is based on a finer-scale regional and recent historical basis rather than global and evolutionary comparisons. In effect, the current study attempts to demonstrate the applicability of the same algorithm within a historical and regional subtext.

Since the landmark publication by Rosenberg *et al.* (2002), the STRUCTURE algorithm has been extended to include three ancestry models, the no-admixture model, the admixture model and the linkage model (section 3.5.5). For all analyses it was assumed that allele frequency is correlated, i.e., there are shared alleles due to admixture, a feature added to STRUCTURE which relaxes the assumption of Hardy-Weinberg equilibrium in the original model.

It is important to note that, by assuming correlated allele frequencies, estimates of K may reflect deviations from random assortment which are not caused by genuine population subdivision, e.g., admixture since population divergence. Thus, because of the presence of such possibilities, the genetic distance between pairs of populations can be over- or under-estimated (Falush *et al.*, 2003). Admixture is probably the main factor to be considered in most of the current study populations, and so its inclusion is an important consideration in the use of the algorithm for the current analyses. The ambiguity surrounding admixture and

population division is the reason why the linkage model is also utilized and assessed. With the linkage model one can also consider correlations in ancestry, i.e., linkage disequilibrium, to provide a more realistic model of historical population genetic processes in human populations.

Although a Bayesian model with several priors, the main focus of STRUCTURE analyses is concentrated on only one parameter, *K*, the estimated number of clusters. Since Pr(*K*) is likely to be very small for values less than the appropriate estimate for *K* (Pritchard *et al.*, 2000), the most appropriate estimate of *K* is the value of *K* at which Pr(*X/K*) plateaus. For all the following analyses this approach was adopted in choosing the correct estimate of *K*.

7.3.1 Inter-population genotypic assignment

The first stage of the analysis was to ascertain if all nine populations could be separately defined, based solely on the autosomal data. The most probable value of *K* for the admixture model was seven (*P(K)* ~ 0.994, Table 7.4a), while for the linkage model nine clusters were most probable (*P(K)* ~ 1, Table 7.4b). Although the linkage model did infer the correct number of clusters in the study sample, there was no unequivocal differentiation into the nine ethnic populations.

From Figures 7.1a and b, three populations could be clearly differentiated, the Han, Hui and Kucong. Initially, the clustering seemed to mirror the summary statistics results, especially the F_{IS} values (section 5.2) with statistically significant F_{IS} values in all three populations; 0.1132 for the Han, 0.4021 for the Hui and -0.0899 for the Kucong. The Han and Hui were distinctive due to their high level of homozygosity, while the Kucong were clearly differentiated as the only population with significant excess heterozygosity.

Miao, Yao and Tibetan individuals displayed less distinct cluster membership. The pattern of multiple cluster probabilities in these three populations was likely to be related to their non-significant F_{IS} values. Furthermore, the fact that the clusters present in the Miao, Yao and Tibetans are also apparent in the Kucong could be ascribed to these populations containing broadly similar southern East Asian ancestral histories But the multiple clustering in the Miao, Yao and Tibetans, unlike the more clearly differentiated Kucong, indicates a history of population admixture.

Table 7.4 Summary of inter-population STRUCTURE analysis.

(a) Admixture Model

K	*lnP(X/K)*	*Var[lnP(X/K)]*	*P(K)*
1	-18036.9	91.4	~ 0
2	-17402	373.2	~ 0
3	-16956	569	~ 0
4	-16618.2	709.5	~ 0
5	-16413.2	833.6	~ 0.006
6	-16439.6	1156.7	~ 0
7	-16408.1	1184	~ 0.994
8	-16587.5	1764	~ 0
9	-16870.5	2588	~ 0

(b) Linkage Model

K	*lnP(X/K)*	*Var[lnP((X/K)]*	*P(K)*
1	-18026.6	79.1	~ 0
2	-17443.7	395.4	~ 0
3	-17022.1	618.8	~ 0
4	-16726.5	848.1	~ 0
5	-16500.6	946.4	~ 0
6	-16595	1380.4	~ 0
7	-16682.6	1660.9	~ 0
8	-17037.3	2663.5	~ 0
9	-16409	1606.3	~ 1
10	-16644.1	2196.8	~ 0
11	-16666.5	2229.1	~ 0
12	-16837.2	2683.5	~ 0

* *lnP(X|K)* is the *ln* of the likelihood of the data given *K*, *Var[lnP(X/K)]* is the variance of the *ln* likelihood and *P(K)* is the probability of *K*.

The history of diverse ancestries in the Bo'an, Dongxiang and Sala would appear to be reflected in the assignment of individuals from these populations to multiple clusters. The clusters to which the Bo'an, Dongxiang and Sala individuals were assigned were, on the whole, distinct from the clusters of the other populations. Application of the linkage model appeared to display a more distinct separation of these three Gansu populations than the admixture model (Figure 7.1b). It is possible that this effect reflected the endogamous nature of the communities since their formation (Wang *et al.*, 2003), a factor also manifested by their statistically significant positive F_{IS} values.

Overall, it is apparent that admixture or endogamy rather than ancestry alone influenced the assignment of individuals into clusters. The STRUCTURE results were broadly reflective of the differing histories and structures of the nine study populations. Therefore the possible presence of substructure in the individual populations needs to be further investigated.

7.3.2 Intra-population genotypic assignment

The next stage of the STRUCTURE analysis was to assess each population separately using the linkage model. The linkage model was selected to take into account finer scale within-population genetic diversity and the relatively low number of loci resolved. Clusters of K =1 to K= 10 were tested for seven of the nine populations. The Tibetans and the Yao were tested with ranges from K=1 to K=4 due to their small sample sizes.

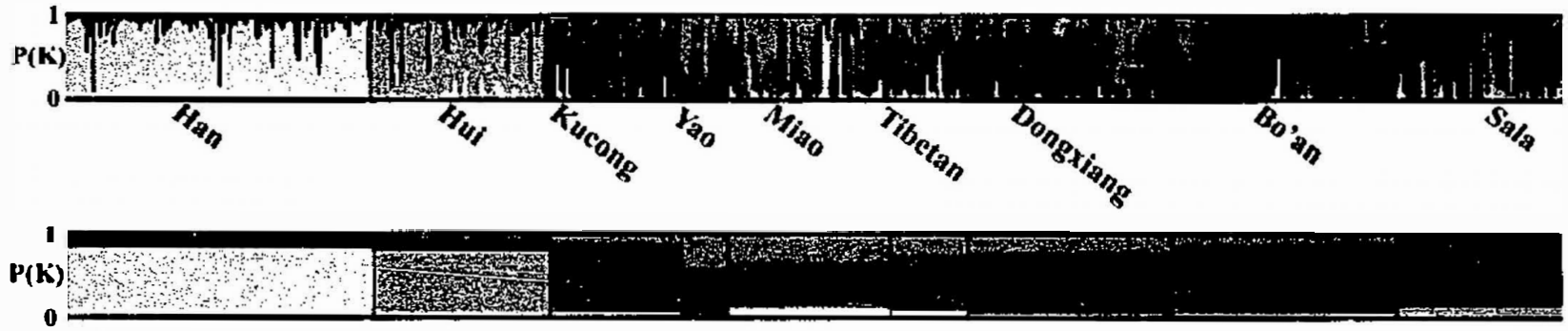
Table 7.5 Summary of intra-population STRUCTURE analysis

Population	*K*	*lnP(X|K)*	*Var[lnP(X|K)]*	*P(K)*
		'Ancient' populations		
Han	3	-2935.3	353.8	~ 1
Kucong	1	-1170.9	41.1	~ 0.97
Miao	2	-1213.6	30.5	~ 0.998
Tibetan*	1	-626.9	22.5	~ 0.817
Yao*	1	-357.5	13.5	~ 0.753
		'Recent' populations		
Bo'an	1	-2576.6	45	~ 1
Dongxiang	1	-2066.1	41.2	~ 1
Hui	3	-1751	267.1	~ 0.998
Sala	4	-2695.2	429.9	~ 1

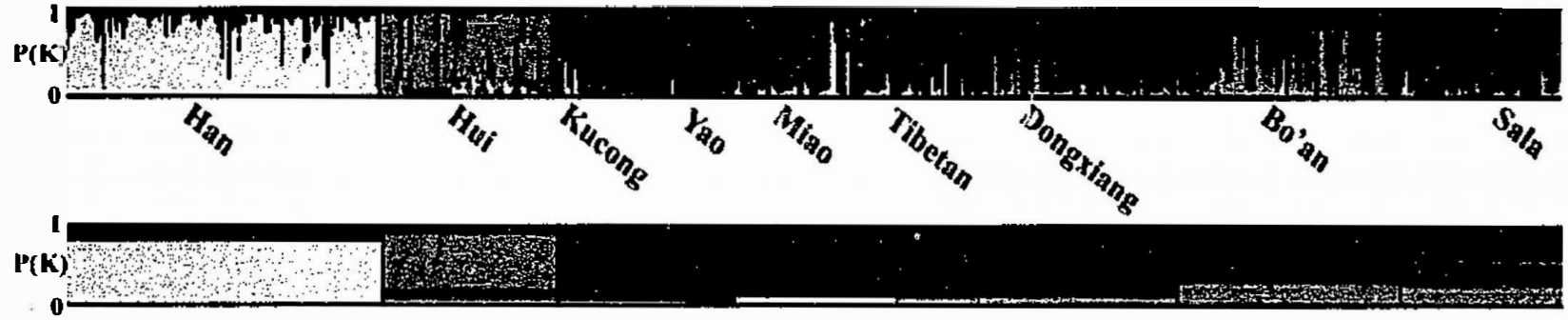
* Population with small sample size. See Table 6.2 for other definitions.

Figure 7.1 Estimated population structure from STRUCTURE analysis of autosomal STRs

(a) Admixture Model K=7



(b) Linkage Model K=9



The top histogram in each model represents the probability of each individual belonging to a particular cluster, with individuals grouped into their original population designations. The bottom histograms in each model represent cluster assignment probabilities in each population.

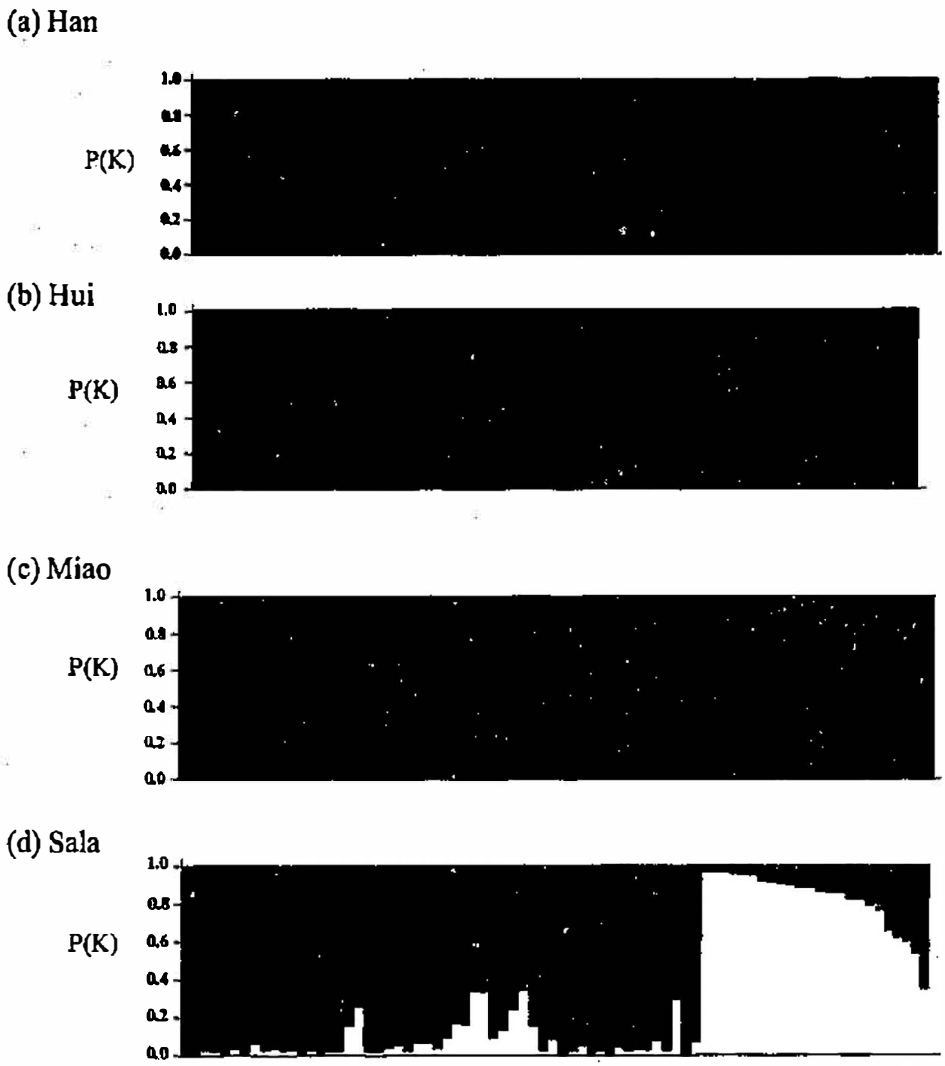
Multiple clustering was present in the Sala, Hui, Han and Miao. The lack of clustering in the Bo'an and Dongxiang is surprising due to strong evidence of the effect of multiple ancestries displayed in the inter-population analysis, but may have resulted from a combination of restricted sample sizes and the restricted number of loci studied. In assessing the results the histories of the populations, their numerical sizes and their internal population dynamics all have to be considered. Each of the study populations had a unique history and was influenced by different factors. In the Hui and Sala, a strong signal due to the dominance of one effect, possibly endogamy, could be inferred. By comparison, multiple effects may have been present in the Dongxiang and Bo'an whose interaction quenched any clear signals.

In the populations where clustering could be inferred, the actual pattern of clustering differed. While the intra-population STRUCTURE analysis showed that the Miao individuals divided into two clusters, the histogram in Figure 7.2(c) shows that all individuals had a 50:50 chance of belonging to one cluster or the other. This pattern is more likely to be a statistical artefact due to the small sample size rather than due to any true population substructure. Alternatively, it could indicate that the Miao community gene pool originated from two distinct sources, which would mirror their previously hypothesised North East Asian paternal ancestries and southern East Asian maternal ancestries.

The clustering patterns in the Han, Hui and Sala were much more distinct and indicated the presence of some form of population substructure. The multiple clusters in the Han matched previous hypotheses in which population subdivision was inferred (Black *et al.*, 2001). Similar scenarios could also be inferred from cluster assignment in the Hui and Sala and would include factors such as the recent formation of each population, coupled with the effect of community endogamy after initial formation. If correct, genetic drift would not have had sufficient generational time to exert a significant effect on the respective gene pools and so multiple clusters, possibly representing different ancestries, remain extant.

The statistical reliability of Bayesian Population assignment within an individual population is not as statistically robust as the inter-population analysis due to smaller sample sizes. It does lend weight to the contention that consideration of

Figure 7.2 Estimated intra-population structure from STRUCTURE analysis of autosomal STRs data in selected populations



The histogram for each population represents the probability of each individual belonging to a particular cluster. For purposes of clarity, individuals were sorted in each histogram according to the membership coefficient (*Q*) for each individual in each cluster.

internal population structure is required when assessing overall human population genetic history. From the results summarized in Tables 7.4 and 7.5 and Figures 7.1 and 7.2, it is clear that ethnic status does not necessarily coincide with genetic heredity. Therefore additional exploration into the definition of ethnicity in China may be required to uncover factors underlying the genetic admixture in these study populations.

Alternatively, the pattern of admixture found in the STRUCTURE analysis could be an artifact associated with the resolution of a small number of loci. It has been shown that the greater the number of microsatellite loci, the clearer the differentiation of populations. For example, in Rosenberg *et al.,* (2002) by increasing the number of loci resolved, a data set encompassing 52 populations from around the globe was more clearly differentiated. In fact, in this study the final number of microsatellite loci resolved totalled 377. The current study had limited financial resources and could not match the statistical power of the Rosenberg study. Nonetheless significant structure is apparent from the limited number of markers resolved. However, the resolution of further markers would further clarify the nature of the population structure and reduce any statistical noise.

The conclusion reached from STRUCTURE analysis is therefore that internal population structure can be inferred from autosomal genetic diversity using a limited number of markers, and it could further be applied to establish whether multiple groupings were present within a population. The causes of genetic substructure could, however, only be inferred in combination with other population genetic sources, e.g., paternal ancestry.

7.4 Bayesian inference of paternal ancestry

The phylogeographic results reported in Chapter 6, most clearly served to differentiate the study populations. As indicated in section 7.1, the final methodology utilized was therefore a full-likelihood Bayesian inference of paternal genetic and demographic parameters under population splitting and growth using the BATWING program (Bayesian Analysis of Trees with Internal Node Generation). BATWING uses a MCMC procedure to generate a sequence of genealogical and population trees, with associated model parameter values, consistent with the genetic

data observed in a sample of individuals. While comprehensive, the BATWING approach does, however, have some limitations, since it does not include factors such as recombination or migration since population formation.

Nonetheless, BATWING is flexible, allowing the use of microsatellite and SNP data, as well as three different coalescent models. Thus in the current study an in-depth investigation of paternal genetic diversity could be undertaken. The three population models applied were the standard coalescent model, a pure exponential growth model, and what in this investigation is termed the beta growth model. A basic representation of these models is shown in Figures 7.3 a-c.

The standard coalescent model (Figure 7.3a) assumes constant size and thus involves the fewest parameters, the major parameter being effective population size (N_e). The exponential growth model (Figure 7.3b) is an extension of the standard coalescent, with pure exponential growth rate and with the term current effective population size (N_c) used to distinguish it from the standard coalescent N parameter. A third population model is available which is more realistic in regard to human history. This model, defined here as the beta growth model (Figure 7.3c), consists of constant growth from the TMRCA until a later time when exponential growth starts, represented by the parameter β. This model has two N parameters, N_a, the ancestral population size, and N_c the effective current population size.

For inter-population studies, a parameter reflecting time of population splitting was included in all three models. In the present study, such population splitting is reflected in a background genealogy constructed from the previously resolved UEP data, thus reflecting the genealogy developed by the Y Chromosome Consortium.

In summary, this section of the investigation was conducted in two parts, inter-population analysis, and haplogroup analysis. By focusing the investigation in this manner it is shown that different aspects of the overall genetic diversity of the study populations can be discerned.

7.4.1 Inter-population analysis

The initial stage of the analysis was an overall survey of the Bayesian phylogenetic relationship between the nine study populations, as an alternative to the

summary-statistic distance methods produced in section 5.3.3. All three population models available were considered.

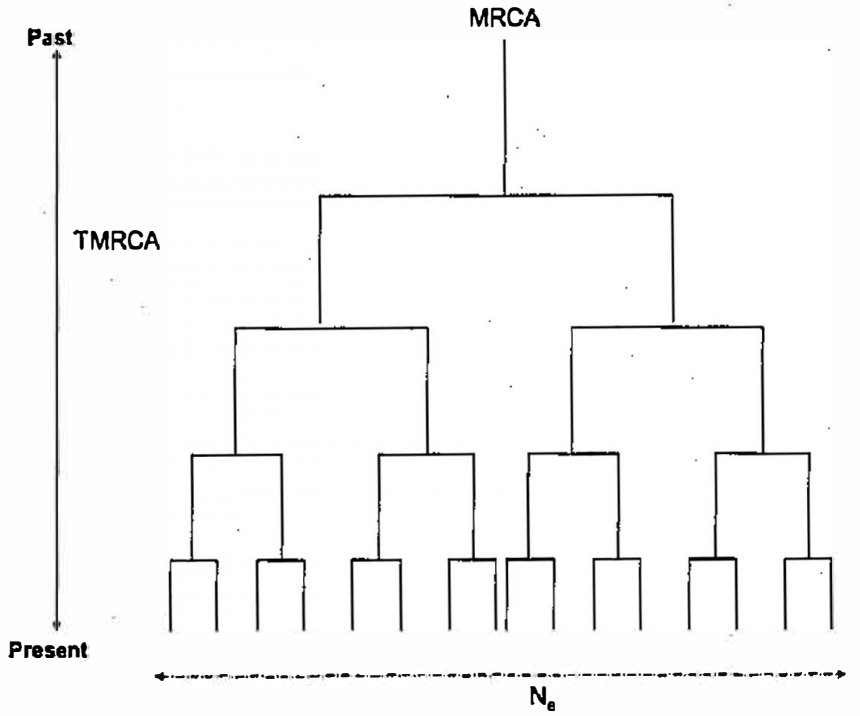
In constructing suitable Bayesian models, the most critical aspect is the selection of priors. After investigation of the literature and an appropriate level of experimentation, most of the priors adopted were based on those defined by Wilson *et al.* (2001); Weale *et al.* (2002); Wilson *et al.* (2003), in which priors were as broad as possible within the limits of estimates of modern human origins.

The prior distribution chosen for the effective population size, N_e, is based on an upper limit from previous estimates of a global effective population size of 10,000. This results in a prior distribution for N_e of *gamma(1.1, 0.0001)*. For the current study, the most recent estimate of the overall effective mutation rate of Y-chromosome microsatellites was utilized in inferring a final prior of *gamma(7, 10000)* for the mutation rate (Zhivotovsky *et al.*, 2004). This method differs from the studies by Weale *et al.* (2002) and Wilson *et al.* (2003), where priors for the mutation rates for each single microsatellite loci were defined.

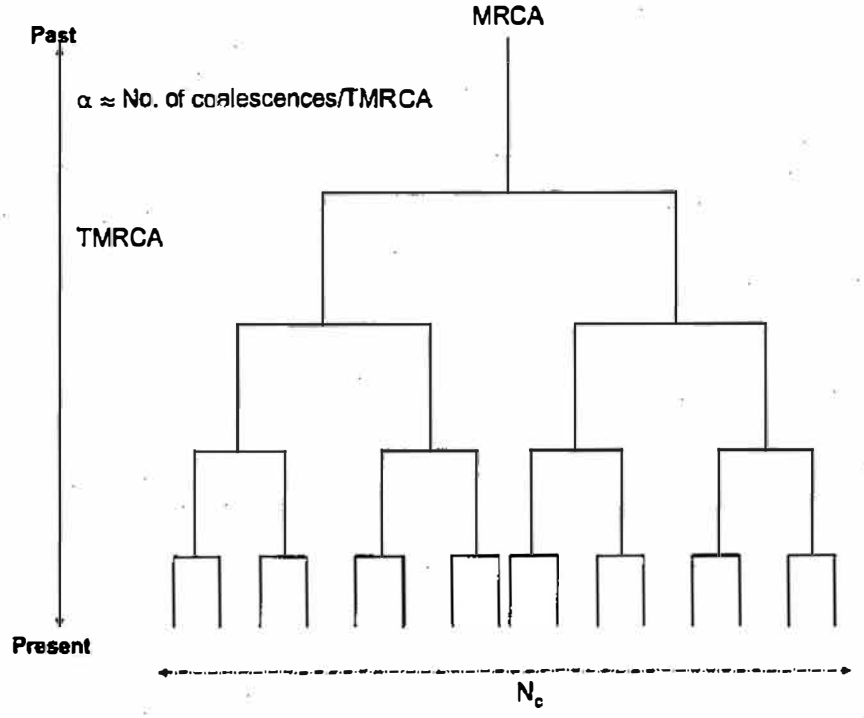
For the inter-population analysis, a prior for the time of the first population split was set. A broad distribution of *gamma(1, 1)* was adopted, to reflect the uncertainty in the ability to reflect the ancestral relationships of the nine populations in this simplified population model. For the exponential growth and beta models a prior for the growth rate, α, had to be defined. A flat broad prior distribution was chosen, as recommended by Weale *et al.* (2003), to reflect the uncertainty in relating

Figure 7.3 Simplified schema of BATWING models utilised in the intra-haplogroup analyses

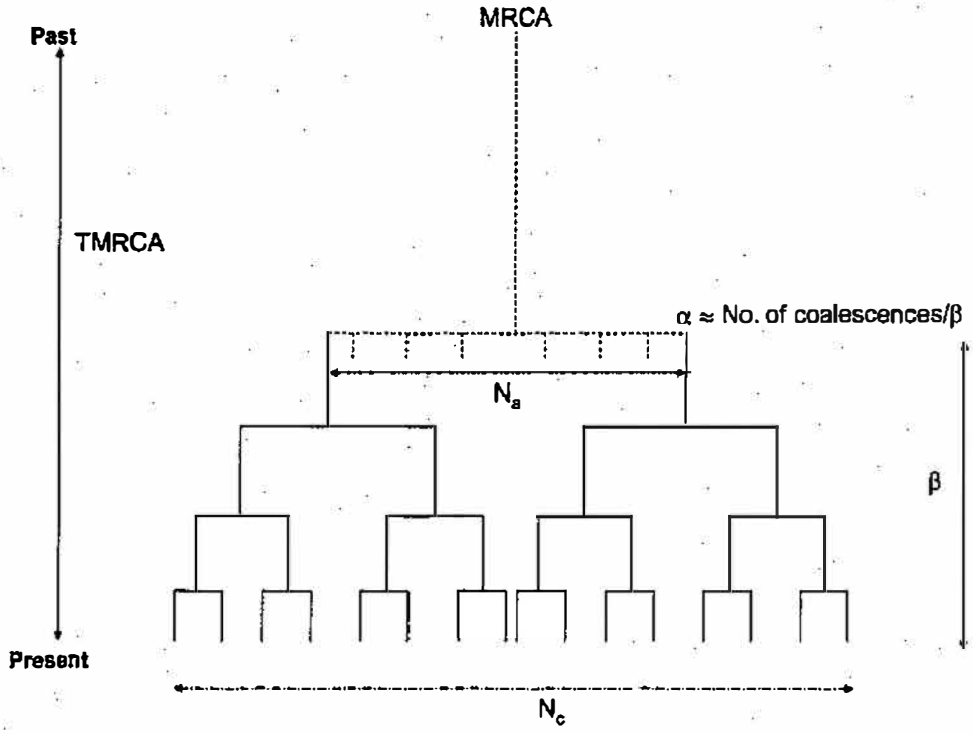
(a) Standard Coalescent



(b) Coalescent with pure exponential growth



(c) Beta growth model



growth in effective population size to growth in real (Census) population size, and it was set as *gamma(1.01, 1)*. A further parameter, β, the start of growth date, had to be defined for the beta model. The BATWING default, flat, uninformative prior was adopted for this parameter, reflecting its uncertain nature for each population.

The data set itself was composed of individuals from all populations. For each individual, eight UEP markers were included plus four microsatellite markers, DYS19, DYS389a, DYS389b and DYS393. DYS388 was not employed due to the discrepancies in its mutational behaviour previously explored in section 6.3.4.1. Unlike studies such as Wilson *et al.* (2001), Weale *et al.* (2002) and Wilson *et al.* (2003) in which UEP data were largely unavailable, the present study included UEP data, and provided a broad background genealogy based on the Y-Chromosome Consortium recommendations. By so doing, uncertainty surrounding the initial inferred ancestry was reduced by limiting the number of possible genealogies in the overall sample space.

The final marginal posterior distributions were estimated from 500,000 BATWING outputs from a total of $2x10^6$ BATWING samples, after discarding a "burn-in" of 200,000. A random sample of 500,000 outputs was made due to computational limitations in the ability to process the data. The final run took approximately 1-2 days on a modest desktop computer (Pentium III 650Mhz) running a Linux OS. Estimates of the posteriors were processed from the final data sample using 'R,, an open-source language and environment for statistical computing and graphics (http://www.r-project.org). A summary of the final results is shown in Table 7.4 and the resultant beta growth model consensus population genealogy is displayed in Figure 7.4.

The results listed in Table 7.6 seem to suggest that the beta model is the most reliable of the three models tested, given previously assigned estimates for the proposed migration out of Africa at around 59,000 YBP, obtained from Y-chromosome studies (Underhill *et al.*, 2001).

Table 7.6 Summary of marginal posterior distributions generated from BATWING inter-population analysis

	Standard Coalescent Model			Exponential Growth Model			Beta Growth Model		
Posterior Quantiles	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
N	17,600	31,000	56,000	21,500	40,000	68,000	2,800	5,500	9,800
μ	0.0005	0.0009	0.0015	0.0002	0.0007	0.0017	0.0004	0.0007	0.0012
α	N/A	N/A	N/A	0.0006	0.001	0.0017	0.004	0.006	0.02
TMRCA	318,000	773,000	204,000	66,300	137,000	323,000	4,800	11,000	28,400
Tg	N/A	N/A	N/A	N/A	N/A	N/A	2,100	4,100	8000

N represents the effective population size (N_e) for the standard model, current population size (N_c) for the exponential model, and ancestral population size (N_o) for the beta model. TMRCA is time to most recent common ancestor in years. For the exponential and beta models, α is growth rate in % per generation. For the beta model, $T_g = GN_o\beta$, where *G* is generation time estimated at 25 years and β is the time of start of exponential growth in coalescent units.
N/A refers to values not applicable to the model.

The standard coalescent model produced a marginal posterior distribution for *N* with a median value of ~31,000 and with a 95% Credible Region (CR) spanning from 17,600 to 56,000, while the *N* posterior distribution of the exponential growth model had a median value of ~40,000, with 95%CR spanning from 21,500 to 68,000. Both models therefore resulted in *N* marginal posteriors encompassing values well above the figure previously estimated at 10,000 (Underhill *et al.*, 2001; Weale *et al.*, 2002). Furthermore, the TMRCA marginal posteriors encompassed values significantly above those estimated by Underhill *et al.*, (2001) of ~60,000 YBP. Therefore both the standard and exponential growth models could be regarded as inadequate for modelling human genetic history.

The beta growth model appeared to be more reflective of human genetic history. The *N* marginal posterior distribution was that of the ancestral population size at the time of start of growth, rather than at TMRCA. This distribution was markedly smaller than the two previous estimates, with a median value of ~5,400 (95% CR 2,800-9,800). The TMRCA had a marginal posterior distribution with median ~11,000 YBP (95% CR 4,800-28,400 YBP), while the 'time of start of

exponential growth' (T_g) marginal posterior distribution had a median value of ~4,000 YBP (95% CR 2,100-8,000 YBP). These are very recent dates in comparison with the previous analyses, and the previously estimated TMRCA of ~60,000 YBP for all lineages out of Africa. In fact the present distributions seem to overlap dates that correspond with the Neolithic population expansion, i.e., ~10,000 YBP. On this basis it could be speculated that most lineages resolved in the current study sample only date back to this event, and that more ancient lineages are either extinct or very rare as an effect of the Neolithic expansion.

This tentative conclusion is strengthened by inspection of the consensus Bayesian phylogeny produced from the analysis, shown in Figure 7.4. All 500,000 possible trees sampled displayed the Kucong as the most ancient population, probably due to their high frequency of F(xK) haplogroups (section 6.3). The Kucong are then most closely connected to the Han. From the Han there are two simultaneous splits. The two clusters produced containing the four 'Recent' and other three 'Ancient' populations respectively.

Direct ancestry cannot be inferred from this tree. Some splits in the consensus phylogeny occur in less than 20% of all trees sampled. This indicates a wide, flat distribution of possible splits indicative of the large amount of uncertainty with regard to any ancestral inference.

Aggregation of the 'Recent' populations is most likely due to the wider range of haplogroups in these populations, and the fact that only these populations contain mutations in markers M45 and/or M17. The much weaker support for the Hui split can be ascribed to their relatively high frequency of the C haplogroup, a more ancient haplogroup, thus adding to a greater degree of uncertainty into the overall genealogical model. In the southern Chinese cluster the Miao and Yao are grouped together with their direct 'ancestor' the Tibetans. This cluster has much less statistical support than the 'Recent' population cluster; from the raw haplogroup analysis the Miao would be expected to cluster more closely with the Han, and the Tibetans with the Yao (section 5.3.2).

Literary sources suggest that the study populations had very different histories, with a broad distribution of ancestries covering much of Eurasia. This is important as the basic assumption of the BATWING population genealogy model

posits a strict binary fission process from an initially panmictic population into a series of isolated sub-populations, and with no subsequent migration. Thus in effect, population admixture is not taken into account. Yet, as has been extensively discussed, admixture forms a very important aspect of the current study. Given the different genetic structure of each of the study populations, the 'one population model fits all' approach taken in the inter-population BATWING analysis may be inadequate to represent their true genetic histories.

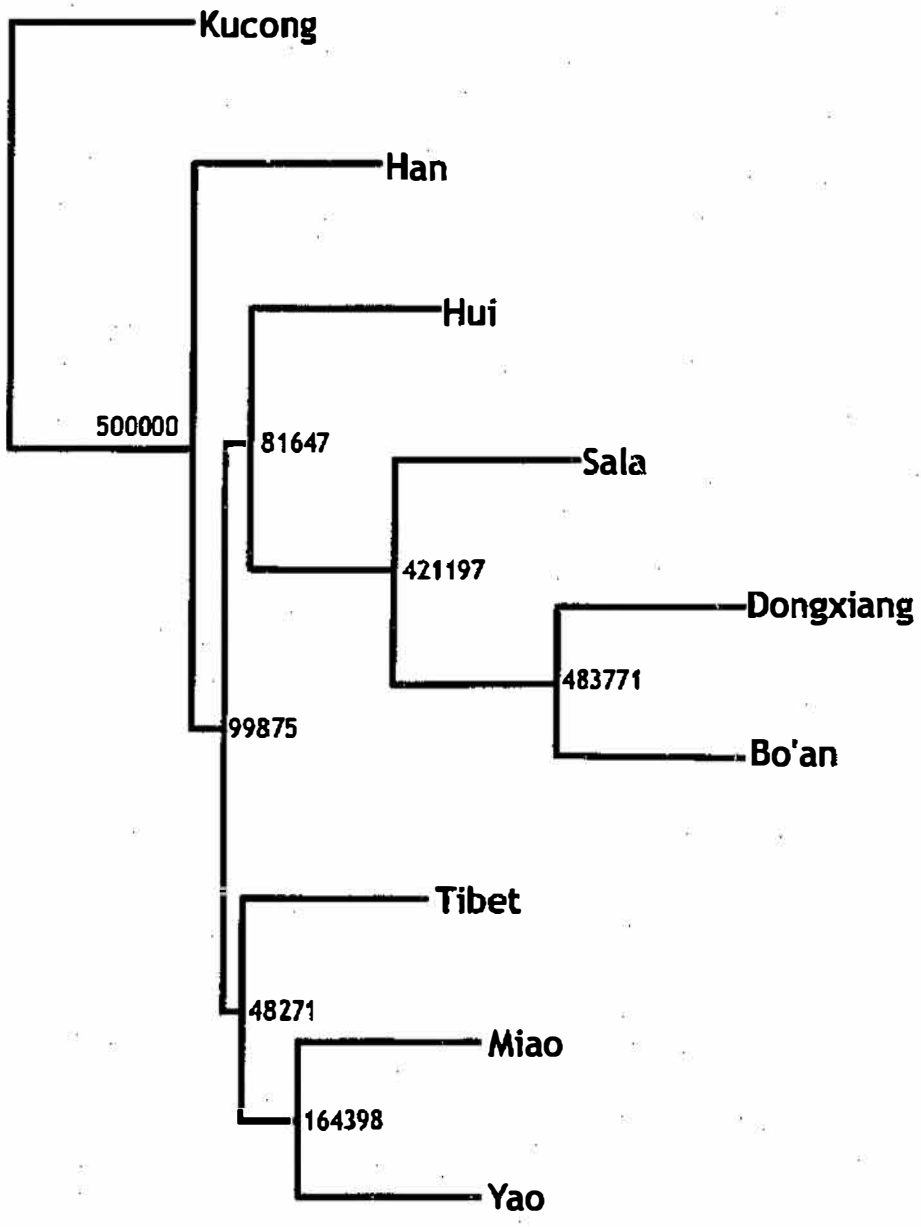
7.4.2 Batwing intra-population analysis

For a more in-depth approach, individual haplogroups within separate populations were investigated in order to reduce uncertainty in the modelling, and thus better resolve the ancestral aspects of the paternal gene pools of each of the study populations. The approach adopted was identical to the previous inter-population analysis (section 7.4.1). This analysis also served as an alternative to the *rho* analysis in section 6.3.5. Therefore key haplogroups within the study populations were investigated to estimate the age of each haplogroup and possibly the age of some of the populations.

This strategy was followed as each population has a wide range of haplogroups, in particular the 'Recent' populations. The key haplogroups chosen represent the most distinctive in each population. Unfortunately, due to their wide spread of haplogroups within a small population size, no statistically reliable BATWING haplogroup analysis was possible for the Hui.

With the exception of those in haplogroups F(xK) and J2, the microsatellite haplotypes analysed were constructed from markers DYS19, DYS388, DYS389a, DYS389b and DYS393. DYS388 was excluded for haplotypes belonging to haplogroups F(xK) and J2 for the reasons given in sections 6.3.4.1 and 7.4.1.

Figure 7.4 Consensus Bayesian phylogenetic population tree from the beta growth BATWING model



The figures indicate the number of trees, from a sample of 500,000, which had the particular population split.

The exponential growth model was not included in any of the analyses because this model resulted in simulations in which equilibrium was not reached, and thus appropriate results could not be resolved. The particular problem with the exponential growth model probably stemmed from the small sample sizes involved.

7.4.2.1 Haplogroup F(xK)

The first step was to explore the structure of the Kucong sample population via analysis of microsatellite haplotypes within the F(xK) haplogroup, which was the most common in the Kucong with a frequency of 0.524. It was on this basis of these samples that a more ancient past was previously hypothesised for the Kucong. The results of the analysis are shown in Table 7.7.

Table 7.7 Summary of marginal posterior distributions produced by BATWING analysis of Kucong Haplogroup F(xK) samples

	Beta Growth			Standard Coalescent		
Posterior Quantile	2.50%	50%	97.50%	2.50%	50%	97.50%
N	15	370	2,000	250	1,000	4,200
μ	0.0002	0.0006	0.0012	0.0002	0.0006	0.001
α	0.003	0.6	3.6	-	-	-
TMRCA	5,500	19,800	78,000	7,400	27,000	112,000
Tg	350	4,700	32,000	-	-	-

Parameters are as previously defined in Table 7.3.
A dash represents values not applicable to the standard coalescent model.

The TMRCA distribution, with a median value of ~20,000-30,000 YBP for both the standard coalescent and beta growth population models, was roughly in keeping with estimates for the arrival of modern *Homo sapiens* into East Asia derived by Underhill *et al.* (2001). The distributions have very broad 95% credible regions, indicative of a large level of uncertainly, probably associated with a combination of small sample size and relatively small microsatellite haplotypes.

The *Tg* distribution has a median of just 4,700. Given the breadth of the credible region of the posterior distribution, this estimate overlaps the period of the Neolithic population expansion at 10,000-8,000 YBP, and indicates a more recent origin for the modern Kucong population than previously suggested (section 6.3.4.2). However, for this conclusion to hold, it is assumed that the beta growth model itself is reflective of the population structure of the Kucong as a whole.

It is believed that, until very recently, the Kucong were composed of small hunter-gather communities, and their population structure may not adequately or appropriately be reflected by the preconditions of the beta growth model. A more constant growth pattern may be a better realisation of the Kucong past. Therfore under the limitations imposed by the BATWING methodology the standard coalescent model, which does not assume a growth rate, may provide a better model of Kucong population history. If this is the case, the Kucong may indeed have a more ancient origin, as was indicated in the inter-population BATWING analysis (section 7.4.1). However, additional demographic, anthropological and genetic studies would be required to resolve whether or not this was the case.

From a genetic perspective, only M89 mutational status was assessed in the Kucong and further sampling and sub-haplogrouping would be desirable in determining their ancestral origins. Combining the BATWING results with the previous summary statistic (Chapter 5) and phylogeographic results (Chapter 6), however, lead to the conclusion that the Kucong do indeed represent an ancient episode of human history in East Asia.

7.4.2.2 Haplogroup J2

A similar BATWING analysis was performed on Han individuals who displayed the J2 haplogroup, a member of the broader F(xK) grouping. For haplogroup J2, only the Han population had an adequate number of samples to enable a statistically reliable analysis to be conducted (Table 7.8).

The Han samples displayed a very different overall posterior genealogy to the Kucong. The most striking results are the broad TMRCA marginal posterior distributions in both the beta (67,300 YBP) and standard coalescent models (median 97,000 YBP),

Table 7.8 Summary of marginal posterior distributions produced by BATWING analysis of Han haplogroup J2 samples.

	Beta Growth			Standard Coalescent		
Posterior Quantile	2.5%	50%	97.5%	2.5%	50%	97.5%
N	70	590	5,000	1,300	4,300	14,400
μ	0.0002	0.0006	0.001	0.0003	0.0006	0.001
α	0.01	0.68	3.7	-	-	-
TMRCA	24,800	67,300	222,000	31,400	97,000	376,000
Tg	3,000	39,400	136,200	-	-	-

and the broad T_g distribution in the beta growth model. The TMRCA estimates are well above the estimate by Underhill *et al.* (2001) of 59,000 YBP, with a 95% CI of 40,000–140,000 YBP. Small sample size may have resulted in a large degree of uncertainty, and thus a wide flat distribution.

As previously explained (section 6.3.1), haplogroups J2 and J are uncommon in East Asia and are thought to have originated in the Middle East. It is therefore possible that these present results reflect the more ancient origins of the haplogroups as a whole, rather than the population history of the Han *per se*.

Another possibility is that the broad distributions may indicate the presence of multiple lineages. From previous phylogeographic and summary statistic results (section 6.3), a significant level of population substructure seemed probable with the degree of admixture of different sub-J2 haplogroups. Thus further definition of these sub-haplogroups would be desirable.

7.4.2.3 Haplogroup O3

Selected samples of haplogroup O3 were tested under the beta growth model. The standard coalescent model was not included in the final analysis, as from the previous results (sections 7.4.1, 7.4.2.1) it was inferred that it was not as representative of human population structure as the beta model, with the possible exception of the Kucong. The results of the analysis are given in Table 7.9.

Table 7.9 Summary of marginal posterior distributions produced by BATWING beta growth model analysis of Han, Miao and Sala haplogroup O3 samples

	Han			Miao			Sala		
Posterior Quantile	2.5%	50.0%	97.5%	2.5%	50.0%	97.5%	2.5%	50.0%	97.5%
N	100	280	900	30	530	2,000	970	2,900	9,000
μ	0.0002	0.0006	0.001	0.0002	0.0006	0.001	0.0002	0.0006	0.001
α	0.03	0.7	3.7	0.006	0.6	3.7	0.007	0.7	3.7
TMRCA	15,300	37,000	105,000	13,000	35,000	114,000	44,000	130,000	440,000
T_g	13,000	29,400	79,200	5,000	12,000	40,000	800	6,000	26,000

Sufficient data were available for analysis of the O3 haplogroup in three populations, the Han, Miao and Sala. The TMRCA and *Tg* Distributions in the Han were similar to those resolved for haplogroup J2, comprising wide flat distributions with medians around 40,000-30,000 YBP. This result did not concur with hypotheses that the O3 haplogroup is a key grouping indicative of the Neolithic population expansion. However, haplogroup O3 has many sub-haplogroups, all of which could be represented in this Han sample. Therefore a higher level of resolution of the UEP markers determining these subgroups would be required in the Han to better model their population structure.

The limitation of the data set was further demonstrated by the results obtained with the Sala. Here the TMRCA posterior was well in excess of any Y-chromosome based estimates of the 'Out of Africa' hypothesis (Underhill *et al.*, 2001), with a very wide flat distribution, a 95% CR of 44,000-440,000 YBP and a median value of 130,000. In contrast, the T_g marginal posterior was more restricted with a 95% CR spanning ~25,000 years and a median value of just 6,000 YBP. These distributions were probably affected by the *N* posterior distribution median value, indicating an ancestral effective population size of 2,900, although the T_g result was commensurate with a Neolithic timescale. The large and somewhat improbable TMRCA results probably result from a combination of the limited levels of haplogroup resolution,

small sample size, and limitations of the BATWING model. Thus interpretation of the T_g distribution must be treated with caution.

In contrast, the Miao haplogroup O3 results seem to be more reliable than those derived for the Han and Sala. The TMRCA posterior distribution for the Miao was similar to that of the Han, but the T_g posterior supported a population growth increase around the Neolithic period, i.e., a median value ~12,000 YBP. The Miao had already been shown to have less Y-chromosome diversity that the Han population (Section 6.3), thus they may contain only a subset of O3 haplogroups, hence the smaller T_g posterior distribution. Overall, the results for the Miao appeared to indicate population expansion during the Neolithic period in East Asia.

7.4.2.4 Haplogroup DE

The haplogroup represented by the YAP insertion, haplogroup DE, has been described as the defining haplogroup of the modern Tibetan population (Torroni *et al.*, 1994; Su *et al.*, 2000), although more recent studies have shown a predominance of this haplogroup in a number of the Yunnan ethnic populations (Deng *et al.*, 2004; Wen *et al.*, 2004b). It has been previously demonstrated that there is a high frequency of the YAP insertion in the Tibetan community, and also in the Yao (Table 6.2 and Figure 6.3a).

Table 7.10 Summary of marginal posterior distributions produced by BATWING beta growth model analysis of Tibetan and Yao haplogroup DE samples

	Tibetan			Yao		
Posterior Quantile	2.5%	50.0%	97.5%	2.5%	50.0%	97.5%
N	55	720	3130	2	35	300
μ	0.0002	0.0006	0.0012	0.001	0.002	0.003
α	0.02	0.7	3.7	0.003	0.5	3.5
TMRCA	11,980	37,500	136,800	900	3,100	10,000
T_g	1,300	6,700	28,000	90	1,500	12,000

As with previous analyses with other haplogroups and populations, the marginal posterior distributions for the Tibetans support the notion of a Neolithic population expansion. The result of a T_g posterior with a median value approaching the estimated Neolithic period of around 10,000-8,000 YBP, in this case around 6,700 YBP, is also found in the Tibetans. Interestingly the TMRCA distribution is centred around 30,000-40,000 YBP (median 37,500 YBP), which is similar to the results obtained with the Han J2 and O3 TMRCA, and the Miao O3 TMRCA distributions. Once again the results suggest ancient migrations into East Asia some 40,000–30,000 YBP, allowing for the uncertainty imposed by the small sample sizes (Underhill *et al.*, 2001).

A more intriguing result is the more recent TMRCA and T_g estimates for the Yao DE haplogroup BATWING results. As indicated in Table 7.10, the Yao TMRCA marginal posterior had a median value of 3,100 YBP, much smaller than in the Tibetans or any other previous intra-population BATWING analysis. The estimated recent origins of the Yao haplogroup D ancestry is further demonstrated by the T_g distribution with a median of 1,500 YBP and a N distribution with a median of just 35. These results indicate that the haplotypic distribution in the Yao is restricted to a limited number of lineages, which possibly is indicative of a recent split from another ancestral population, probably of proto-Tibetan origins.

An alternative hypothesis is that, as with the Kucong, the beta growth model may not adequately represent the population history of the Yao. Unlike the eight other study populations the μ posterior for the Yao has a much smaller range, with a median value of 0.002. This compares with the values of 0.0006-0.0007 demonstrated by the prior and marginal posterior distributions of other populations in the study, which themselves are consistent with the effective population mutation rate estimate reported by Zhivotovsky *et al.*, (2004). To represent the Yao data in a beta growth model, the mutation rate had to be set much lower than this estimate, and as a result the model itself may not be suitable for the task of assessing their history. For this reason, a standard coalescent population model was tested for the Yao haplogroup DE haplotypes (Table 7.11).

Table 7.11 Summary of marginal posterior distributions produced by BATWING standard coalescent model analysis of the Yao haplogroup DE samples

	Yao		
Posterior Quantile	2.5%	50%	97.5%
N	45	200	770
μ	0.001	0.002	0.003
TMRCA	1,300	4,500	15,000

The results of the standard coalescent model analysis showed a larger TMRCA posterior median value of 4,500 YBP, but a similar μ marginal posterior to that of the beta model analysis. While wider and larger than the beta growth TMRCA posterior, the standard coalescent TMRCA posterior seemed to indicate a more recent date than any other TMRCA marginal posterior from previous analyses. The fact the μ marginal posterior again displayed a high mutation rate, probably indicating that there were only a few repeat differences between the samples and therefore they represented a restricted ancestry. However, further anthropological and pedigree studies on the Yao community sampled would be required to explore this hypothesis.

7.4.2.5 Haplogroup R1a1

The final set of BATWING intra-population analyses was based on haplogroup R1a1 in the Bo'an, Dongxiang and Sala. As previously discussed, this haplogroup was only found in individuals in the 'Recent' populations (Table 6.2), and it is thought to represent the Middle Eastern and Central Asian paternal ancestries of these populations.

All three sample sets displayed wide flat TMRCA marginal posterior distributions, with median values for the Bo'an, Dongxiang and Sala of ~40,000, ~50,000 and ~70,000 YBP respectively. These estimates are well outside the estimates by Underhill *et al.*, (2001) of 30,000-20,000 YBP for this haplogroup, although the distributions do encompass these values near the edges of the 95% CRs.

Table 7.12 Summary of marginal posterior distributions produced by BATWING beta growth model analysis of Bo'an, Dongxiang and Sala haplogroup R1a1 samples

	Bo'an			Dongxiang			Sala		
Posterior Quantile	2.5%	50.0%	97.5%	2.5%	50.0%	97.5%	2.5%	50.0%	97.5%
N	80	740	3,500	240	860	3,200	12	1,200	7,300
μ	0.0002	0.0006	0.001	0.0002	0.0006	0.001	0.0002	0.0006	0.001
α	0.003	0.67	3.7	0.03	0.7	3.7	0.001	0.65	3.7
TMRCA	12,400	38,100	141,300	16,700	47,300	161,200	21,000	66,100	250,000
T_g	600	6,400	40,400	2,700	9,800	31,400	450	13,600	130,000

As with previous estimates, the findings were probably caused by the large degree of uncertainty associated with small sample sizes. Meanwhile, the T_g posteriors showed the now familiar pattern of 95% CRs that indicated a signal of population growth starting in the Neolithic. Haplogroup R1a1 has its origins in Central Asia, but is common throughout the Middle East and into Europe. An investigation of R1a1 haplotypes in Armenia using BATWING seemed to show a signal of population growth that arose in the Levant (Weale *et al.*, 2001). The T_g marginal posterior distributions in the Bo'an, Dongxiang and Sala may similarly represent a signal of eastward population growth, stemming from the Levantine Neolithic expansion of ancestral populations out of the Middle East.

7.5 Summary

The use of three stochastic likelihood-based methods provided further insights into the genetic ancestry and structure of the study populations. Significant population substructure was demonstrated with the MIGRATE-N and STRUCTURE analyses, while paternal ancestries stemming from both the East Asian and Levantine Neolithic have been proposed. With MIGRATE-N a basic difference between the relatively homogenous maternal genetic diversity and extensive paternal genetic diversity was demonstrated.

To further unravel the intricacies of the structure of modern human populations, additional research is needed accompanied by the development of models which more realistically reflect the population genetics of human societies. Initial approaches to modelling variables, such as consanguinity (Overall and Nichols, 2001) and selection (Coop and Griffiths, 2004), in coalescent models have been made. However the developments remain quite preliminary and, as yet, they are incapable of encompassing the complex structures present in the different study populations.

Irrespective of whether a summary statistic or a stochastic likelihood-based analysis is adopted, the central question is does the method match the specific question under investigation? To answer the broader question of the cause and effect of the population structures revealed in the current study, the genetic evidence needs to be inferred in conjunction with historical, anthropological and demographic evidence. A cross-disciplinary overview of the human population history of China incorporating the results of the present study is therefore presented in Chapter 8.

CHAPTER 8

DISCUSSION AND CONCLUSIONS

8.1 Introduction

The results presented in Chapters 5-7 demonstrated that, by assessing the bi- and uniparental gene pools of the nine study populations, different aspects of their histories could be revealed. As part of the Discussion, it will be demonstrated that the broad scope of the current study, by combining demographic, historical and anthropological evidence with multi-system genetic data, has uncovered complex genetic heritages in each of the nine Chinese ethnic populations.

To present a comprehensive overview of the findings of the current investigation, the Discussion will be structured into three major sections. The first part will focus on the structure and ancestry of the nine study populations as revealed by the results of the present investigation. This will then be followed by a general consideration as to the applicability and reliability of the population genetic methods utilised. Finally, overall conclusions as to the findings of the study will be presented.

8.2 The Han

Given their majority status within PR China it is appropriate that exploration of the constituent study populations starts with the origins of the Han. In order to place the results resolved from this study into context, an overview of aspects of Han history and identity are required. Definition of the term Han itself has changed through history, and so the changing definition of ethnicity and identity in China is a major theme of this Discussion.

Many major studies into the genetic diversity of the Han have focused on various aspects of the origin and nature of human prehistory in East Asia. As comprehensively discussed in section 2.3, the most crucial aspect of these studies is the effect of the Neolithic expansion and its influence on modern Chinese genetic diversity. As in previous studies (Su *et al.*, 1999; Su *et al.*, 2000; Yao *et al.*, 2002a; Wen *et al.*, 2004a), the results resolved from the Han samples in the current study displayed strong evidence for the existence of a Neolithic population expansion. This was demonstrated via the dominance of the Y-chr O haplogroups (Table 6.2) and mtDNA haplogroup D (Table 6.1) in the Han sample population resident in Liaoning Province. The wide diversity of Y-chr STR haplotypes within the Han provided additional evidence of a rapid population expansion (section 7.4.2.3).

However, a different aspect of the genetic structure of the Han of Liaoning was revealed by assessing autosomal genetic diversity, i.e., the possibility of population substructure. In addition, the presence of the Y-chr haplogroups C, K*(xO,P), and especially J2 at low frequency, suggest the possibility of recent admixture.

To further explore the nature and origins of the disjoint Liaoning Han gene pool, it is necessary to consider the history of the Han peoples in Liaoning. The area that now forms the southern portion of Liaoning Province was for many centuries the border between the Chinese Empire and the 'barbarian' hordes, the Jurchen (Fairbank and Reishauer, 1990). Through contact with the Chinese Empire the Jurchen tribes learned of Chinese culture, resulting in the formation of a Man (Manchu) nationality and eventually the establishment of the Qing dynasty in 1644 AD (Table 4.1).

It was only after the early years of the Qing Dynasty that mass migration and settlement of Liaoning by Han communities was recorded. The Han migration into present-day Liaoning resulted from the recruitment of Imperial peasants by the Qing Dynasty to bolster its own power. Within a generation, Han immigrants had outnumbered the local Man inhabitants and thereafter they became the dominant ethnic population of Liaoning (Campbell *et al.*, 2002).

Crucially, after the initial settlement of Liaoning, some Han migrants and their descendants changed their family names to Man surnames and thus altered their perceived ethnic status. Conversely, during the late stages of the Qing Dynasty many people with Man family names changed to a Han surname. Campbell *et al.* (2002) concluded that Manchu and Han ethnicity in late Imperial Liaoning was essentially fluid and that many factors other than heredity were important in determining individual ethnicity. It has been shown that the number of males switching from Han to Man names was at its peak during the middle periods of the Qing dynasty (1700-1800 AD), but quickly declined as the power of the dynasty began to wane (Campbell *et al.*, 2002). On this basis Campbell *et al.* (2002) concluded that membership of a specific ethnic group in Liaoning during the Qing period seemed to centre more on economic and political position rather than ancestral heritage.

More recent events could also have affected the genetic structure of the current Han population of Liaoning. One example is that the area now known as

Liaoning (formerly a part of Manchuria) was included in the puppet state of Manzhouguo (Manchuko), which was under direct Japanese rule from 1932-1945 (Fairbank and Reischauer, 1990). Manzhouguo served as a centre for heavy industry with its rich natural resources being used to fuel the Japanese military machine, and with Pu Yi, the last Qing Emperor, installed as its nominal ruler.

After the establishment of PR China in 1949, Liaoning became a leading heavy industrial region, with parts of the province enjoying Special Economic Status. As a designated Special Economic Zone, Liaoning attracted many economic migrants brought in by the state-run industries that dominated heavy industry in the country. However, since the economic reforms of the 1980s and 1990s, which involved a reduction of state-owned heavy industry, the economic strength of Liaoning has been greatly reduced (Becker, 2000). Nevertheless, the influx of economic migrants over a 40-50 year period could have served as a further major source of genetic diversity within the Han majority population of the province.

With this history of fluid ethnic affiliation and mass migration, it is difficult to regard the Han of Liaoning, let alone the entire Han *minzu* as a homogenous ethnic community, and in terms of genetic diversity the results of the present study are in keeping with this concept. The analysis of autosomal diversity showed excess homozygosity amongst the Han sample population, presumptive evidence of internal population sub-structure.

From the Y-chromosome and mtDNA haplogroup analyses there also was evidence of the late admixture of different Han groups representing different migration events, and possible admixture with Man communities. While Y-chr O haplogroups and mtDNA haplogroup D predominated in the Han, other haplogroups were present in lower frequencies. The Y-chromosome haplogroup C, which is uncommon in East Asia but common in North Asian populations, could be representative of possible Man ancestry (Table 6.2). Furthermore, the presence of the mtDNA haplogroups B and R*, which are more commonly seen in southern China, suggested immigration into Liaoning from more southern regions of the country (Kivisild *et al.*, 2002; Yao *et al.*, 2002a; Wen *et al.*, 2004b).

However, the results obtained cannot be regarded as conclusive and further demographic and pedigree studies would be required to resolve the detailed structure

and ancestry of this Han community and reconcile the nature of admixture in the present population.

8.3 Tibet and the Tibetans

Inferring genetic history from current genetic diversity is similarly complex in the non-Han ethnic minorities. Inferring genetic history and structure for the Tibetan population has proved especially difficult, since it involves interpretation over multiple time periods. Furthermore, the investigation of Tibetan history and origins is more controversial due to the current unsettled nature of the political situation of Tibet and the Tibetan peoples.

So, who are the Tibetans? The currently accepted hypothesis dates the origin of the modern Tibetan population back to the East Asian Neolithic population expansion (Qian *et al.*, 2000; Su *et al.*, 2000; Deng *et al.*, 2004; Wen *et al.*, 2004b). From this event there was the formation and subsequent migration of peoples into Northwest China termed the 'proto-Tibeto-Burman' migrations, and via these migrations the ancestral population of the Tibetans and other populations of Northwest China were formed.

This migration theory does not, however, explain the dominance of Y haplogroup D in the Tibetans (and its absence elsewhere in China). One hypothesis is that the presence of Y-haplogroup D in the Tibetans is due to a Central Asian contribution to the proto-Tibetan gene pool approximately 6,000 YBP (Su *et al.*, 1999; Su *et al.*, 2000; Deng *et al.*, 2004). Since that period haplogroup D lineages in Central Asia have largely been replaced by other lineages, such as R1a1, which resulted from later population admixture due to the nomadic nature of communities in the area, especially after domestication of the horse. By comparison, in more remote areas such as the Tibetan Plateau, the geographical isolation of the population could have resulted in the persistence of the haplogroup D lineages.

The Central Asian contribution theory was originally advanced by Qian *et al.* (2000) in their major study into Tibetan Y-chromosome diversity, and in which they cited Altheide and Hammer (1997) who advocated an Asian origin for the Y-Alu Polymorphism. Since this report, the development of UEP analysis has allowed the determination of a more complex haplogroup family, haplogroup DE, of which

haplogroup D is an exclusively Asian subgroup. From this broader definition of haplogroup DE, it has been concluded that the YAP insertion is of ancient African, and not a more recent Asian, origin (Underhill, 2000; Underhill, 2004). Furthermore, to date haplogroup D has only been found at significant frequencies in Tibetans (Qian *et al.*, 2000; Wen *et al.*, 2004b), Japanese (Hammer and Horai, 1995; Tajima *et al.*, 2004), Andaman Islanders (Underhill, 2004) and some ethnic communities of Yunnan province, PR China (Qian *et al.*, 2000; Deng *et al.*, 2004; Wen *et al.*, 2004b). According to Underhill (2004), this distribution of the YAP insertion can best be regarded as a relic distribution of ancient pre-Neolithic remnant lineages.

From the molecular data analysed as part of the present study, either of these two main Tibetan origin scenarios could be inferred. BATWING analysis of haplogroup D STR haplotypes in the Tibetan sample population for the current study seemed to show that both the Central Asian and pre-Neolithic origin hypotheses were possible. The TMRCA of ~37,500 YBP (Table 7.10) was supportive of the proposed first migrations into China around 40,000 YBP, albeit with a very wide 95% CR ranging from ~12,000 YBP to ~137,000 YBP (Underhill *et al.*, 2001; Underhill, 2004). The TMRCA estimate was also supported by *Rho* analysis, which suggested a TMRCA of ~30,800 YBP (Table 6.3). However, for the BATWING analysis, the time of start of population growth was timed at around ~7,000 YBP (95%CR ~1,300-28,000 YBP), thereby allowing for the proposed ancient Central Asian contribution. Further experimental evidence would be required to decide definitively between the alternative hypotheses.

Additional evidence for the origins of the Tibetans can, however, be acquired by reference to various geographical, climatological and anthropological aspects of the Tibetan Plateau. Firstly, the model of post-Neolithic migration into Tibet would involve agricultural peoples from relatively low elevations moving on to the Tibetan Plateau, which is in part hyper-arid and frigid and is situated at the highest permanently settled altitude on Earth, with consequent selection pressures in terms of hypoxic stress (Beall *et al.*, 2004). Acclimatisation to this sparse high altitude environment in which pastoralism, and not agriculture, is better adapted could thus have been a problem for Neolithic migrants. Secondly, the Tibetan plateau is composed of many sharply separated geographical regions differing in altitude and

environment, from high mountain passes to low foothills and valleys (Aldenderfer and Yinong, 2004). Extensive archaeological investigations have uncovered artefacts from these different environments representative of many different communities and cultures (Aldenderfer and Yinong, 2004). This long and diverse record of human occupation in Tibet dates from before the Neolithic expansion to the fall of the Tibetan Kingdom in the 9th century AD, a kingdom which at its peak had spheres of influence ranging from modern Sichuan and Qinghai Provinces in PR China to Nepal and India (Du and Yip, 1993; Ebrey, 1999; Adshead, 2000).

Tibet is now most commonly recognised as that area incorporated with the Tibet Autonomous Region. However, the Tibetan Plateau itself extends much further, reaching into Qinghai Province, Sichuan Province and the northern tip of Yunnan Province. The Plateau also covers regions along the Himalayas in India, Nepal and Bhutan and this wider definition of Tibet has been termed 'ethnographic' Tibet (Aldenderfer and Yinong, 2004). In effect, the larger area represented by 'ethnographic Tibet' encompasses many more ethnic groups than the official Tibetan *minzu*, and it illustrates the need to account for greater cultural and genetic diversity when attempting to assess the history of the Tibetan peoples.

Given the geographic, climatological, anthropological and archaeological complexities of the Tibetan plateau, it seems pertinent to consider a newly proposed hypothesis on Tibetan ancestry. One possibility is an origin stemming from both a demic expansion that accompanied the Neolithic expansion and an acculturation mechanism, i.e., that the modern Tibetan population has resulted from an uptake of cultural and technological aspects of the Neolithic without a concomitant uptake or transfer of genes (Aldenderfer and Yinong, 2004). Under this scenario, molecular data would indicate a Tibetan origin from the admixture of pre- and post-Neolithic groups, and lineages with direct ancient origins could easily have been lost and/or admixed with those of post-Neolithic migrants. Thus, depending on the community sampled, genetic evidence of more ancient origins could or would not be forthcoming. In the present study population, a BATWING estimate of 7,000 YBP was reached for the start of population growth, which could be inferred as indicative of an admixture event, rather than direct evidence of any purely migrant origin.

By combining past genetic studies, the present DNA-based investigation, and the geographical and archaeological evidence, the joint demic expansion /acculturation hypothesis appears to be the more probable. Only further archaeological and anthropological evidence will, however, serve to confirm or reject this hypothesis.

8.4 Yunnan populations: the Miao, Yao and Kucong

The Kucong, Miao and Yao are just three of at least 25 ethnic populations that have been identified in Yunnan Province, described as the most 'ethnically diverse' province in PR China (Qian *et al.*, 2001). The ethnic diversity of the Yunnan population is due to the geographical position of Yunnan between Tibet, Southeast Asia and greater mainland China. The area now know as Yunnan was not part of any Chinese state until it was conquered by the Mongol (Yuan) Dynasty during the 13th and 14th centuries AD. Before that time, two kingdoms had been established, firstly the Nanzhao kingdom established by the Bai during the seventh century AD and then the kingdom of Dali in the 10th century (Du and Yip, 1993).

Most importantly, Yunnan was the centre of a southwest silk route, and so formed a base for the trading of many goods, especially horses and precious metals (Yang, 2004). It is inferred from Chinese historical texts that many different trade routes were established during the period of the two kingdoms, the major route being from Sichuan, and through Yunnan to Burma, Assam and India (Yang, 2004). These kingdoms became allies of the Chinese Emperor against a common foe, the Tibetan Kingdom, which resulted in an increase of trade with the Chinese Empire and multiple migrations southwards from northern and central China into Yunnan (Du and Yip, 1993; Wen *et al.*, 2004a).

The present-day ethnic diversity of Yunnan is therefore not surprising, given the regional history of near constant migration from within and beyond China. It is against this migrational background that the genetic diversity of the three Yunnan study populations, the Miao, Yao and Kucong is most appropriately assessed.

8.4.1 The Miao

In approaching the interpretation of the genetic diversity of the Miao sample population, the difficulty of defining the Miao culturally and historically quickly becomes apparent. An anthropological study of the Miao suggested difficulties in classifying them as a single ethnic group (Diamond, 1995). In fact, early uses of the term Miao (which literally translates as 'sprouts', 'seedling' and in some cases 'weeds') referred to most of the disparate indigenous peoples across Southwest China rather than to a single ethnic group (Diamond, 1995). The Miao can be split into three broad groups, the Hmong (located in Yunnan), the Hmu (in Guizhou) and the Xioob (located in Guizhou and Hunan). However, cultural and linguistic differences are observed within these three groupings, mostly at village level (Diamond, 1995).

There also are at least three major Miao languages, each of which in turn is divided into multiple dialects. Furthermore, the Hmong of Yunnan are not necessarily connected with the Hmong peoples of Southeast Asia, and the terms Miao and Hmong are used interchangeably (Diamond, 1995).

In summary, the Miao *minzu* comprises a diverse range of communities. From the results of the present study, intra-community genetic diversity was apparent with a marked differentiation in maternal and paternal ancestries between a mostly northern male component and a southern female component. A majority of the Miao paternal lineages were of haplogroup O3, i.e., the haplogroup thought to be the major marker of the Neolithic expansion from northern China. However, the mtDNA haplogroup distribution of the Miao was dominated by haplogroups B, F and R*, which are most common in southern China and Southeast Asia and are rare in northern China. The results therefore seem to indicate that the Miao study population had a paternal ancestry sourced from migration(s) from northern China, and that these male migrants then married local southern Miao females.

The dominance of the Y-haplogroup O, and thus a similar Yellow River Basin origin to the Han, is in agreement with a hypothesis derived from Chinese anthropological texts stating that the Miao originated from the Yellow River region as part of a southern migration. The Miao community in the present study had a strong Northeast Asian male ancestral component and so had fewer links with the

Southeast Asian Hmong communities, despite the latter being referred to as 'Miao' in Chinese literature (Diamond, 1995).

Many Western scholars also interchangeably use the terms Miao and Hmong. This confusion between Miao and Hmong, especially since they are classified in the same linguistic grouping, is a typical example of the confusion frequently encountered in the identification of ethnic communities in Yunnan Province, and PR China in general.

8.4.2 The Yao

As with the Miao, definition of the Yao ethnic group is fraught with problems. The Yao *minzu* is spread throughout six different Provinces of China, mainly in the mountainous regions which extend across the south of the country. However, there also are Yao living in the neighbouring countries of Myanmar, Vietnam and Thailand. As alluded to in section 4.1.1.8, the Yao have a long history of continued migration (Lipman, 1995). Linguistically, the Yao language is composed of many dialects and it is categorised in the Hmong-Mien grouping along with the Miao dialects. This is despite the fact that the Yao and Miao languages are mutually unintelligible (Litzinger, 1995).

There are over thirty different 'subgroups' or *zhixi* in the official Yao *minzu* structure, and there are over three hundred different terms for the Yao in the Chinese language (Litzinger, 1995). The cultural diversity of the Yao *minzu* is demonstrated by the worship of separate deities in different Yao communities and by community-specific beliefs in different mythological origins and religious festivities, none of which were allegedly known to other Yao communities before establishment of the Yao *minzu* in 1949 (Litzinger, 1995). Therefore investigations into Yao genetic diversity must be approached at the local community level.

The Yao sample in the current study appears to represents a particular Yao community with two major ancestral links, Tibeto-Burman and Southeast Asian. The distribution of Y-haplogroups was very similar to the Tibetans, where the majority of individuals belonged to haplogroup D with a minority of O haplogroups. By comparison, the mtDNA haplogroups resolved in the Yao displayed a mixture of northern and southern origins, with haplogroups D, B, R and F all present. As there

were only a limited number of Yao samples available for these analyses, further sampling would be required to strengthen these results. Nonetheless, it is inferred that the Yao have a common ancestry with the Tibetan study population.

The suggestion that the Yao represent a southern migration from a larger proto-Tibetan population was further strengthened by BATWING analysis (Section 7.4.2.4), where the age of the Yao lineages was estimated to be much younger than that of the Tibetans (Tables 7.10 and 7.11). With a restricted distribution of haplotypes, these Yao lineages could date from the establishment of the Tibetan Kingdom, as the Yao of Yunnan were mainly resident in the northern tip of Yunnan, an area under the influence of the Tibetan Kingdom during the height of its power.

The strong presence of Y-haplogroup DE in the Yao did suggest a Tibeto-Burman background, thus demonstrating a linguistic phylogeny linking the Miao and the Yao, but it does not necessarily equate with genetic ancestry. The Miao had a definite paternal Northeast Asian ancestry, while the paternal ancestry of the Yao was Tibeto-Burman. In fact the languages and dialects within the Hmong-Mien language group are very diverse, and it seems that they may have been grouped together because of their geographical proximity rather than linguistic similarity.

The Yao sample population investigated represents just one of many culturally diverse communities within the Yao *minzu*. Therefore it is possible that by investigating other Yao communities many different genetic ancestries could be uncovered, with further investigation into the definition of Yao nationality required to further determine the origins of these peoples.

8.4.3 The Kucong

The Kucong were the least known of the study populations. As discussed in section 4.4, few anthropological and historical sources are available on these people. Their hunter-gatherer lifestyle in the remote Ailaoshan Mountains of Yunnan Province ensured isolation from other neighbouring communities. After the establishment of PR China many Kucong were relocated into agricultural villages and towns (Xiaotong, 1991). The Kucong speak the Lahu-shi language of the Lahu, a population which is not only a Chinese *minzu*, but also is present in Myanmar,

Laos, Thailand and Vietnam (Xiaotong, 1991). In fact, Lahu itself is a very broad term covering many subgroups (Du and Yip, 1993).

Because of this linguistic affiliation, the Kucong were placed under the Lahu *minzu*. However, in all other cultural aspects the Kucong have little in common with the Lahu and they have probably adopted the Lahu language only quite recently. Thus what little documentary information is available on the Kucong provides a confused picture of multiple classifications (Xiaotong, 1991). The current study provides the first genetic evidence of the identity and ancestry of the Kucong and therefore allows some biological information on their identity.

Firstly, the maternal genetic diversity of the Kucong proved to be significantly different from the other study populations. Estimated mtDNA haplogrouping showed a predominant Southeast Asian component of ancestry, with a majority of individuals displaying B, F or R* haplogroups (Table 6.1). Northeast Asian haplogroups C and A were also resolved from some Kucong individuals. However, the basic mtHV1 phylogenetic and summary statistical analysis showed a sharp difference between the Kucong and the other study populations (Section 5.4 and Figure 5.6). This latter result is due to the presence of local, and possibly community-specific, Kucong HV1 sequence polymorphisms absent from any of the other study populations. The sharp difference may be indicative of older isolated mtDNA haplogroup lineages in the Kucong, thus providing further confirmation of a more ancient gene pool. Any definite conclusion, however, would need further investigation of mtDNA HV2 and mtDNA coding region polymorphisms.

Inferring the ancestral origins of the Kucong is somewhat more reliable from the Y-chromosome data. The most obvious signal is the predominant frequency of Y-haplogroup F*(xK), indicated by the presence of M89C-T polymorphism, but absence of any other polymorphism tested for (Table 6.2). It is believed that individuals carrying the M89C-T polymorphism formed part of the first wave of migrants into East Asia. Outside East Asia this haplogroup, and its non-M9 sub-haplogroups, are very common in Europe (Semino *et al.*, 2000), Central Asia (Wells *et al.*, 2001), and the Indian subcontinent (Quintana-Murci *et al.*, 2004).

The presence of haplogroup F*(xK) in this isolated community in Yunnan province suggests either recent migration from outside China or much earlier origins

stemming from the first migrations into East Asia. Given the known hunter-gatherer lifestyle and long-term isolation of the Kucong, the latter scenario seems the more likely. There have been suggestions that the M89 haplogroups in India and Central Asia stem from expansions from East Asia into Central Asia and the Indian Subcontinent (Wells *et al.*, 2001). If this hypothesis is correct, the Kucong may represent remnants of ancient source populations for this expansion. There were also some Kucong individuals carrying another ancient remnant Y-haplogroup, haplogroup DE, which reinforces the ancient origin scenario.

Finally, additional evidence that the Kucong harbour ancient lineages was demonstrated by the restricted pattern of Y-STR haplotype diversity (section 5.3.2), which indicated haplotypes that contained mainly small DYS19 alleles of 13 and 14 repeats. This restricted haplotype pool and the presence of smaller DYS19 alleles not found in the other study populations was interpreted as support for ancient paternal origins.

By combining the Y-chromosome evidence with mtDNA-derived data, the ancient origin scenario does seem much more probable than the recent migration hypothesis. At the very least, the Kucong represent an opportunity to further determine the origins and nature of the establishment of *Homo sapiens* in East Asia. However, once again it must be emphasized that further demographic and anthropological studies are required to confirm if this inference of the Kucong as an ancient indigenous East Asian population is correct.

Any proposed anthropological and demographic study into the population may need to proceed within a short time-frame, due to aspects of Kucong history tentatively inferred from autosomal diversity. Unlike the other study populations, the Kucong sample contained excess heterozygotes, as evidenced by the significant negative F_{IS} value (Table 5.1), which is most likely to have caused their clear separation from other populations in STRUCTURE analysis (section 7.3). The reason for this finding was unclear given the lack of information on their history and demography. However, one possibility that could cause heterozygote excess is recent population amalgamation (Hedrick, 2000). Relocation of the Kucong into villages and towns after the establishment of PR China (Xiaotong, 1991) may well have resulted in the agglomeration of previously isolated Kucong sub-communities.

The policy of relocation demonstrates the need for further demographic and anthropological investigations into the Kucong to elucidate their population structure and history. If sub-community agglomeration has indeed occurred, it may already have resulted in the loss of some of the unique genetic heritage of the Kucong. Therefore the sooner additional studies can be commenced the better, lest a possible link to an ancient past may be irretrievably lost.

8.5 The Silk Road and the arrival of Islam in China

From the ancient past, the focus of the discussion now shifts to the more recent historical events. The Silk Road, which proved to be an important source of genetic diversity in China , was a series of trading routes spanning from Xi'an through Northwest China and Central Asia to Constantinople and far as Italy. According to Chinese historians, the Silk Road originated in the 2nd century BC when the Emperor Wu Han sent a court official, Zhang Qian, to seek allies in the western regions (Northwest China) to enhance the security of the Han dynasty. The mission ultimately failed, but Zhang Qiang did gain knowledge and information about the lands farther afield in Central Asia, and presumably the populations of Central Asia also first learned of Han China, with reciprocal trade links eventually established.

Besides Western goods, establishment of the Silk Road also led to the arrival of foreign settlers in China. For example, in 61 AD some 10,000 Turkic families from the Central Asian city-states of Samarkand and Bokhara were recorded as settling in Xi'an (Du and Yip, 1993). Other recorded entries into China include the influx of Persian refugees from what is now modern Iran after the fall of the Sassanid Empire to the Arabs in 652 AD.

The arrival of Islam into China had a more lasting effect. The history of Chinese Muslims dates back to the Tang Dynasty (618-907 AD), when followers of the Islamic faith, predominantly men, were reported to have variously entered China as soldiers, merchants and political emissaries from Central Asia, Arabia and Persia (Leslie 1986; Wong and Dajani 1988, Du and Yip 1993; Lipman 1997, Gladney 1996, 1998). At least some of these men settled in China and married local Han

women, thus founding the Chinese Muslim communities (Wong and Dajani, 1988; Du and Yip, 1993).

From these founding settlers, the history of Islamic communities to the founding of PR China has often been one of assimilation with the majority Han. Prior to the fall of the Yuan (Mongol) Dynasty in 1368 AD, Muslim communities were largely separated from the majority Han, although during the Yuan period Muslims were entrusted with important responsibilities and powers, primarily acting as middlemen between the Mongol overlords and the majority Chinese population (Leslie, 1986). From the formation of the Ming Dynasty, a program of assimilation of Chinese Muslims into the local Han populations was undertaken, and by the time the Qing Dynasty came to power this process of assimilation was complete (Leslie, 1986; Wong and Dajani, 1988). Assimilation was reversed by the Qing, with all Muslims listed as a separate population group, following the enactment of special laws promulgated for this purpose (Leslie, 1986). It was not until the founding of PR China in 1949 that separate Muslim ethnic groups were, however, formally recognised.

The history of diverse ancestries, assimilation and then recognition as separate ethnic groups is strongly reflected by the multi-layered patterns of genetic diversity uncovered in the four Muslim communities investigated in the present study. By investigating maternal and paternal gene pools and autosomal diversity it was found that different episodes in the histories of these communities could be revealed from the molecular genetic data.

In all four Muslim communities, the Bo'an, Dongxiang, Sala and Hui, mtDNA lineages largely consisted of haplogroups of East Asian origin. This was best demonstrated by the haplogrouping described in section 6.2 in which haplogroups D, A, B, C, F, Z were resolved, all of which have been described as being of an East Asian origin (Kivisild *et al.*, 2002; Yao *et al.*, 2002a; Wen *et al.*, 2004b). The use of MDS and phylogenetic techniques further demonstrated that, from a maternal perspective, it was difficult to separate these Muslim 'Recent' populations from the other 'Ancient' populations (Figures 5.5, 5.6), which conforms with the historical evidence of a local Han female origin for the Muslim populations.

In contrast, investigation of the patterns of Y-chromosome molecular diversity in the four Muslim *minzu* demonstrated a wide range of paternal ancestries, possibly spanning the Middle East, Central Asia, Northern Asia and East Asia. The distribution of these lineages in each population also showed that each individual population displayed different genetic histories which were not necessarily concomitant with their linguistic classification.

The Sala were the population sample that most clearly displayed this linguistic-genetic dichotomy. As discussed in section 4.1.1.7, the Sala were believed to be of largely Turkic origin with further admixture of Mongolians, and the present-day language spoken by the Sala reflects this proposed community history since it is a dialect that is effectively Turkic with some Mongolian and Han words (Du and Yip, 1993). The genetic evidence resolved in the current study also attested to this view of history with individuals belonging to the Y-chr DE, C, J2, R1a1 and O haplogroups, which represent a variety of ancestries (Table 6.2). At the same time, over half of the males in the Sala sample population belonged to either Y-Haplogroup O* or O3, possibly indicating a stronger local Han contribution than first hypothesised.

Overall, despite the population using a Turkic language, only a minority of the male individuals tested had evidence of Turkic ancestry. Therefore, either the Sala population was originally formed from a small community of Turkic paternal ancestry which Mongolian and/or Han Chinese males joined and adopted the Turkic language, or the language could have been introduced into the founder population during their migratory history.

Language and genetic ancestry non-correspondence can also be demonstrated in the neighbouring Bo'an and Dongxiang. Both of these populations speak a Mongolian dialect but, as with the Sala, the paternal genetic evidence suggested a much broader genetic ancestral pool. Indeed, the small Bo'an community displayed the greatest Y-haplogroup diversity of each of the study populations and contained individuals with all the haplogroup lineages tested (Table 6.2). Once again, the present-day language may either have been introduced or adopted during migration, or it was the language and background of the founder population into which

individuals with Han and Central Asian background entered and adopted the language, in this case Mongolian.

In contrast, the northern Mandarin-speaking Hui of Liaoning exhibit the lowest frequency of the Y-haplogroup(s) O of the four populations (Table 6.2). For the Hui of Liaoning, use of the local language may be associated with their strong merchant history. The Hui *minzu* itself is a widely devolved nationality found in almost all of the provinces of China, a situation which is testament to their mercantile history and Hui Muslims still figure strongly in trading in present-day China (Ma, 1996; Gladney, 1998). The Hui of Liaoning are no exception and the lengthy tradition of Liaoning as a major Chinese industrial centre may have been a major factor in attracting the community to settle in the region.

While initial female Han assimilation into the Hui may have occurred, the Y-chromosome and autosomal genetic diversity of the Hui of Liaoning suggest a history of genetic isolation and endogamy. Previous studies established the occurence of consanguineous marriages and more general endogamy within the Hui population of Liaoning via random sample collection and the investigation of extended family pedigrees (Black, 2000; Black *et al.*, 2001).

Reanalysis of the autosomal data, especially using STRUCTURE analysis, confirmed this autosomal homogeneity in the Hui (Table 5.1, Section 7.3). Therefore despite their diverse founding ancestries, and historical evidence of forced assimilation, the Hui have largely practised intra-community marriage resulting in the preservation of diverse Y-chromosome lineages but with a restricted autosomal gene pool.

The restricted patterns of autosomal genetic diversity in the three Muslim Gansu populations, the Bo'an, Dongxiang and Sala, also seem to demonstrate some degree of endogamy. Each community exhibited significant heterozygote deficiency, indicated by positive F_{IS} values (Table 5.1). However, these values were smaller than the levels found in the Hui, and for the Dongxiang the F_{IS} was only 0.01. Unlike the Hui, no detailed Bo'an, Dongxiang and Sala pedigrees were available to assess direct evidence of consanguineous marriage and endogamy. Nonetheless, previous studies utilising historical and demographic evidence showed that the populations were largely endogamous (Baric, 2000; Wang *et al.*, 2003). As these

populations have disjoint maternal and paternal gene pools, further anthropological investigations into their structures are required to positively confirm the extent of consanguinity and the endogamy of the populations as a whole. In summary, by analysing three complementary genetic systems, different aspects of the population history and structure of the four 'Recent' Muslim populations, the Bo'an, Dongxiang, Hui and Sala were revealed.

8.6 The concept of ethnicity in China

Analysis of the genetic structure and history of the nine study sample populations demonstrated a clear difference between the concepts of ethnicity and genetic ancestry in PR China. The eight sample populations classified as ethnic *minzu*, the Bo'an, Dongxiang, Han, Hui, Miao, Sala, Tibetans and Yao, showed a wide array of genetic structures and ancestries, thus questioning the concept of the *minzu* as a useful population label for genetic studies. Initially the single non-*minzu* sample population, the Kucong, displayed a structure and ancestry that seemed to better fit the concept of a *minzu* than the eight officially recognised populations surveyed. However, the *minzu* concept is not solely anthropological in nature, and can be considered as the current end-result of a long Chinese history of philosophical and political definitions of ethnicity.

Definitions of ethnicity and ethnic identification in China have a long and complex history. The concept of the 'Middle Kingdom' was central to Chinese thinking, and it probably remains so to the present day (Fairbank and Reischauer, 1990). Throughout Chinese history the ruling elite viewed China as the centre of the world, with all other states in the 'barbarian' fringe. Foreign 'barbarians' were labelled *wayji* (outside barbarians), whereas minorities who lived within the Middle Kingdom were identified as *neiyi* (inside barbarians) (Dikötter, 1992). This practice was in accordance with the Confucian concept of *zheng ming* (rectification of names), whereby labelling and categorisation restored and maintained order and thereby ensure that all was well with the world (Leslie, 1986). By so doing, the ruling dynasties could create the impression of firm control over their dominions.

An example of the concept and practice of categorisation was that defined by the Mongol dynasty (1279-1368 AD), whereby the population of Imperial China was

divided into four groupings: The Mongols; *Semu* ('the coloured eyes', i.e. Western and Central Asians); *Hanren* ('Han people', who included Northern Chinese, Khitans, Jurchens and Koreans), and *Nanren* ('southerners') (Dikötter, 1992). After the establishment of the Peoples Republic of China in 1949, the Communist Party adopted similar approaches to exert its power through a centralised pattern of Government, resulting in the establishment of 56 official *minzu*. The definition of a *minzu* by the Government of PR China consisted of key markers such as common territory, language and economy, and was largely based on the Stalinist concept of nationality (Gladney, 1998).

The official *minzu* system of PR China is little different from the older concepts of ethnic groups in that it seems to create a sense of order out of the complex ethnography of the country, providing the government at least the appearance of central control over the populace. Recent investigations of Chinese genetic diversity, including the present study, have primarily relied on the official *minzu* classification to classify their population samples (Ding *et al.*, 2000; Su *et al.*, 2000; Yao *et al.*, 2002b; Yao and Zhang, 2002; Wang *et al.*, 2003). However, as the results of the present study have demonstrated, the genetic structure and ancestry of the different officially recognized ethnic groups of PR China are much more varied than their *minzu* classifications would appear to indicate.

As demonstrated by the flexible ethnicity adopted by the Han and Man communities of Liaoning, the concept of ethnicity in China is not necessarily directly correlated with genetic ancestry. In fact, after the process of formalisation of the 55 minority *minzu* following the establishment of the Peoples Republic, many Han community groups in Yunnan who lived in close association with ethnic minorities were able to adopt minority category labels in order to benefit from the political and personal advantages that accrued from ethnic minority membership (Tapp, 2002). Another example of ethnic 'flexibility' is provided by the Yao and Miao *minzu*, both of which are composed of many sub-communities as previously noted. Before the establishment of the *minzu* these sub-communities had little, if any, contact with each other because of their different dialects and cultures. As such, from a demographic and a genetic perspective, they could be considered as different ethnic groups.

The problem of Chinese ethnic identification is most apparent in Yunnan Province, which is the most ethnically diverse region of the country (Du and Yip, 1993). Within Yunnan, different communities have lived in varying degrees of complex inter-relationships for at least 2000 years (Peters, 2001). The result is that in some cases, cultural attributes, like religious beliefs and costume, cannot be regarded as accurate indicators of ethnic and genetic ancestry. The Blang, an Austro-Asiatic Mon-Khmer-speaking people who live in the valleys of the Xishuangbanna region of Yunnan provide a good example of this cultural borrowing. The Blang have adopted the culture of the predominant ethnic community in the area, the Tai Lue. They wear Tai Lue costume, speak the Tai Lue language as well as their own, practice the Tai Lue religion of Theravada Buddhism, and they have even learned wet-rice agriculture from the Tai Lue. Therefore, at first appearance, it is very difficult to differentiate between a Blang and a Tai Lue (Peters, 2001).

It is probably more appropriate to consider the concept of *minzu* as a convenient political construct as much as it is an ethnological identifier. As part of the *minzu* system, minority groups are allowed a limited degree of administrative autonomy within various Autonomous Counties, Prefectures and Regions. Furthermore most of the 55 minority *minzu* have not been subject to the One Child Certificate programme, and as a result their numbers have increased disproportionately since its inception in 1979 and with less evidence of biased secondary sex ratios (Bittles and Chew, 1998, Festini and de Martino, 2004). The creation of official *minzu* has also served to increase a sense strong national identity in some ethnic groups, for example, among the Yao, who have evolved an increased notion of a Yao nationality and identity among their disparate communities (Litzinger, 1995; Tapp, 2002), with publications by Yao scholars of a unified Yao history where other historical, anthropological and linguistic studies question this concept (Litzinger, 1995). A similar growth in nationalism has been reported amongst the Hui (Gladney, 1998) and the Muslim Uygur in Xinjiang province (Gladney, 2002). Thus the recent growth in nationalism based on ethnicity is perhaps starting to defeat the original purpose of the *minzu*, i.e., to enhance central control over the disparate peoples of PR China.

These different political, demographic and historical aspects of the concept of ethnicity in China demonstrate that a cautious approach to population structure is required, in order to accurately assess the ancestry and genetic diversity of its many ethnic populations. It therefore is recommended that any future studies into human genetic diversity in China should initially seek to define the detailed structure and possible sub-structure of proposed ethnic study populations, using historical, anthropological and demographic sources. This procedure should ideally be considered on a community by community basis, and reliance solely on the official *minzu* category as an accurate indicator of genetic ancestry should be avoided.

8.7 Population genetics and population modelling

The final aspect of the present study consisted of the application of a range of population genetic techniques to assess the wider issues of genetic diversity in PR China. An appropriate understanding of how best to classify ethnicity and population structure in PR China is particularly pertinent both in the selection of population genetic methods, and the subsequent interpretation of results. For the present study, a broad range of methods and models were utilised in order to interpret the raw molecular data resolved from the sample populations. It was found that each method, from simple summary statistics to the more complex likelihood-based stochastic modelling, described different aspects of the history and structure of the study populations.

The more comprehensive analyses, like AMOVA and phylogenetic construction, highlighted important aspects of inter-population genetic relationships, including similar mtDNA diversity but dissimilar Y-chromosome relationships. However, the basic classical population genetic techniques, such as levels of heterozygosity/gene diversity proved most useful in assessing the internal structure of each population.

Recent studies on Chinese genetic diversity have mainly concentrated on the overall pre-historic genetic ancestry of various ethnic populations, using Y-chromosome or mtDNA polymorphism haplogrouping and network analysis (Ding *et al.*, 2000; Yao *et al.*, 2002b; Wen *et al.*, 2004b), with little emphasis on internal population structure. In the current study, the simpler task of assessing the levels of

heterozygosity in the different Chinese populations proved effective in assessing more recent historical effects. It also revealed the need to explore fine-scale structure within ethnic groups in order to place the ancestral origins of each population, revealed by mtDNA and Y-chromosome UEP haplogrouping, into the historical context.

The developing field of statistical phylogeny has been a major focus of this study, however the results from the three techniques adopted, MIGRATE-N, STRUCTURE and BATWING, were mixed. MIGRATE-N (section 7.2) while more complex and at times unstable, gave results which most closely reflected those inferred from the basic summary statistical analyses (sections 5.3 and 5.4). In effect they indicated that the 'Recent' populations had greater autosomal diversity than the 'Ancient' populations, that the level of mtDNA diversity was similar in each population, and that the Y-chromosome data provided little evidence of male admixture. However, MIGRATE-N was not reliable for analysis of the mtDNA HV1 sequences, because of the increased computational complexity involved in properly modelling this unusually unstable sequence in the large data set resolved in the present study.

The STRUCTURE algorithm (section 7.3), although employed for only 10 variable autosomal STR markers, was useful in demonstrating the structural differences between the study populations. The results obtained were, however, somewhat vague and this method is probably better applied to the large-scale studies into overall human genetic evolution for which it was originally designed.

The Bayesian BATWING method did prove effective in generating estimates of TMRCA and the times of foundation of each population, but a limitation was the simplified population models on which the analyses were based. While adequate on a global evolutionary basis, the three available population models did not include important and relevant demographic variables known to have influenced the histories of the study populations, such as migration and admixture after their initial formation of the population. Furthermore, the wide confidence region of the inferred marginal posterior distributions seemed to engender some doubt as to the reliability of the final estimates of the parameters included in the models. In a previous study it was reported that a 95%CR lower limit of 2 years was produced for T_g for a particular

population, and on this basis is was concluded that BATWING analysis produced erroneous results (Zerjal *et al.*, 2002). These observations may simply reflect the significant uncertainty surrounding the parameter estimates, i.e., uncertainty as to the initial mutation rate estimates and the appropriateness of the inferred population model. Bayesian modelling encompasses the degree of uncertainty surrounding a parameter estimate, and thus can result in wide and sometimes seemingly implausible marginal posterior distribution confidence regions. The results obtained may not necessarily be incorrect, but may simply reflect the lack of prior information available on the model parameters. As such they arguably could be considered to be more realistic than summary statistic based methods, such as the rho statistic (section 6.3.4.2).

It is concluded that due to the infancy of the development of likelihood-based stochastic techniques, and limitations in available computational power, caution should be exercised in adopting techniques such as MIGRATE-N, STRUCTURE and BATWING for applied molecular anthropological investigation. The population models utilised are broad evolutionary models, for example, the exponential growth models utilised in MIGRATE and BATWING, where only a limited number of demographic variables and coalescent models are explored. Actual population structures would involve a large number of different processes including migration, admixture, endogamy, consanguinity, bottlenecks and population growth, and in addition these processes may only occur during specific time-periods (Harding and McVean 2004, Knowles, 2004). As yet, an all-encompassing likelihood-based stochastic technique of this nature remains to be developed (Knowles, 2004). It is also important to realise that in populations with complex histories, specific events or processes may fail to leave an identifiable genealogical signal (Hudson, 1990; Knowles, 2004). That is, it may not be possible to reveal all possible events and processes that could have resulted in the current genetic diversity of a study population from molecular data alone. Therefore, even if feasible, a broad-based model that incorporates multiple possible variables may prove not to be effective.

A more practicable approach would be to gain a greater understanding of the structure and history of the population in question from anthropological and demographic sources and then construct a population model, possibly at a kinship or

a pedigree level, on which analyses of molecular data can be based. With this type of approach, it should be possible to infer which demographic processes are most likely to affect the current genetic structure, and thus develop and utilise specific population genetic models applicable to that particular scenario. This approach was illustrated by Chaix *et al.* (2004) in their investigation of the recent genetic history of the Vlax Roma for whom extensive pedigree data were available, with more defined priors and/or models for MIGRATE-N and BATWING explored.

From the results of the current study the application of similar approaches to the vastly more diverse Chinese *minzu* were useful in defining pre-historic ancestral origins, but they proved not to be helpful in defining more recent historical aspects. The lack of reliable anthropological, historical or demographic data for some of the population samples was a major contributory factor to this shortcoming. As discussed in preceding sections, what little information was available indicated that the classification of a sample population as one specific ethnic grouping was probably misleading. In effect, the *minzu* classification was too broad a category for the accurate inference of molecular genetic ancestry, and the definition of credible sub-communities would be required to derive more accurate estimates of parameters such as TMRCA and T_g.

The main focus of any future human genetic diversity study in China should therefore not be on the merits of what a particular population genetic model can or cannot do, but on the level of anthropological and demographic detail required to enable the development of assumed population structure. In short, a stronger emphasis is needed on the initial experimental planning and design.

CHAPTER 9

RECOMMENDATIONS FOR FUTURE STUDIES

9.1 Introduction

The scope of the present investigation has, at least in part, demonstrated the wide range of disciplines needed in attempting to adequately depict the genetic heritage of PR China. Concentrating on just one aspect of Chinese genetic population structure, whether Y-chromosome diversity (Su *et al.*, 2000; Deng *et al.*, 2004) or mtDNA diversity (Yao *et al.*, 2002b), will not necessarily result in an adequate or realistic definition of the overall history and structure of the populations in question. In fact, even studies which use both Y-chromosome and mtDNA information e.g., Wen *et al.* (2004b), may not necessarily produce an accurate depiction of either population structure or history, given the size and complexity of the populations living in PR China.

9.2 Consequences and recommendations for future studies

To effectively explore the genetic history of Chinese ethnic populations, the present study confirms that, in addition to Y-chromosome and mtDNA diversity, an analysis of autosomal diversity should be conducted. The analysis of autosomal markers helped to uncover aspects of recent population effects, such as admixture and endogamy, which the interpretation of male and female ancestral origins and of genetic diversity in general would not have revealed.

Simply relying on genomic data alone will, however, not necessarily be successful in revealing all aspects of the genetic structure of ethnic groups in PR China. While the initial focus of the present study was on molecular and statistical techniques utilised to investigate genetic ancestry and structure, with ongoing investigation the importance of the initial experimental design and small sample size became an increasingly significant issue. It is the interpretation of population history from a range of sources, including anthropological, demographic, archaeological, geographical and historical information, which is crucial in any molecular anthropological study, and inadequacies in the initial definition of candidate populations can lead to misleading conclusions from genomic data.

The changing definition of ethnicity and national identity through the course of Chinese history illustrates the need to interpret Chinese ethnic population history with due caution, and over-reliance on *minzu* classifications could prove costly.

Even with the more sophisticated coalescent-based population genetic methods under development, greater in-depth knowledge of population structure and sub-structure is required for future studies into genetic diversity within and between Chinese ethnic populations.

Basic population model inferences, however, may have to be altered in attempting to obtain better models of human population structure. The models commonly used in BATWING, for example, assume simple coalescent models from constant growth, simple exponential growth, and a bottleneck effect (beta growth). An alternative approach, such as the meta-population model proposed by Harding and McVean (2004), could usefully be incorporated and tested for internal population structure investigations. This would especially be the case where autosomal markers are used and alternate processes that do not appear to follow the 'Recent African Origin' hypothesis have been found (section 2.23).

Another possible approach to explore fine-scale ethnic population structure in PR China would be to investigate the concepts of kin structure and kin-structured migration (Fix, 1997, 2004), with kin-structured migration a particularly relevant topic in the present study. The classical population genetic view of migration is based on abstract models, such as the stepping stone model (Hedrick, 2000). The effect of kin-structured migration is to increase local gene frequency differences. This contrasts with classical migration models, in which migration serves as a force that reduces differentiation (Fix, 2004).

The concepts of kin structure and kin-structured migration have been investigated in molecular ecological and classical ecological studies of many different species, from humpback whales (Valsecchi *et al.*, 2002) to social insects including ants, bees and wasps (Hammond and Keller, 2004) and our closest living primate relative the chimpanzee (Morin *et al.*, 1994; Gagneux *et al.*, 1999). However, the term kin structure is infrequently used or discussed with respect to the molecular anthropology of humans. This omission is all the more surprising since kinship is widely accepted as a significant, fundamental factor throughout human history, and it manifestly served to structure many of the populations investigated in the present study, more especially the Muslim Bo'an, Dongxiang, Hui and Sala.

One example where the concepts and effects of kin structure were explored was an investigation into the genetic consequences of traditional marriage customs in Pakistani and Indian communities in the U.K. (Overall and Nichols, 2001; Overall *et al.*, 2002; Overall *et al.*, 2003). These studies involved the collection of detailed pedigree data and more general family histories, coupled with the generation of genomic data. The results demonstrated that the effects of fine-scale processes such as consanguinity and endogamous marriage were significant.

Importantly from a public health perspective, it was concluded that estimates of disease gene frequencies could only be accurate if made jointly with estimates of population substructure and consanguinity (Overall *et al.*, 2003). The investigation thus led to the researchers resolving many different substructures in their study communities, and also to a redefinition of the relative risk of recessive disorders and forensic DNA typing approaches in these groups.

A similar approach to the genetic study of Chinese ethnic communities is recommended, especially amongst the southern Chinese populations resident in Yunnan Province where ethnic and cultural diversity is maximal. It is recognised that this proposed approach would be tantamount to reviewing the entire Chinese concept of *minzu*, and the overt political implications of seeking to investigate and review ethnic identity in PR China would require both caution and tact. However, in time the end-results of such a revision should be seen to have merited the means.

In conclusion, with major projects such as Hapmap (http://www.hapmap.org) and the Genographic Project (http://www.nationalgeographic.com/genographic) now underway, it is recommended that the complexities of internal population structure and the fluidity of ethnicity, at least in PR China, be considered if the aims of these projects are to be fully realised.

REFERENCES

Abdo, Z., Crandall, K. A., & Joyce, P. (2004). Evaluating the performance of likelihood methods for detecting population structure and migration. *Mol Ecol*, **13** (4), 837-851.

Adshead, S. A. M. (2000). *China in world history* (3rd ed.). Basingstoke U.K., Macmillan Publishing.

Arredi, B., Poloni, E.S., Paracchini S, Zerjal, T., *et al.*, (2004). A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet*, **75** (2), 338-345.

Aldenderfer, M., & Yinong, Z. (2004). The prehistory of the Tibetean plateau to the seventh century A.D. : perspectives and research from China and the West since 1950. *J. World Prehis*, **18**, 1-55.

Altheide, T. K., & Hammer, M. F., (1997). Evidence for a possible Asian origin of YAP+ Y chromosomes. *Am J Hum Genet*, **61** (2), 462-466.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., *et al.* (1981). Sequence and organization of the human mitochondrial genome. *Nature*, **290** (5806), 457-465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., *et al.* (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, **23** (2), 147.

Ayub, Q., Mansoor, A., Ismail, M., Khaliq, S., *et al.* (2003). Reconstruction of human evolutionary tree using polymorphic autosomal microsatellites. *Am J Phys Anthropol*, **122** (3), 259-268.

Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*, **16** (1), 37-48.

Bandelt, H. J., Forster, P., Sykes, B. C., & Richards, M. B. (1995). Mitochondrial portraits of human populations using median networks. *Genetics*, **141** (2), 743-753.

Bandelt, H. J., Macaulay, V., & Richards, M. (2000). Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogenet Evol*, **16** (1), 8-28.

Baric, T. (2000). Genetic Diversity in the Bo'an, Salar, and Dongxiang: Co-resident Muslim Populations in Gansu Province, P.R. China, BSc. Honours Thesis, Edith Cowan University, Perth, Western Australia.

Beall, C. M., Song, K., Elston, R. C., & Goldstein, M. C. (2004). Higher offspring survival among Tibetan women with high oxygen saturation genotypes residing at 4,000 m. *Proc Natl Acad Sci U S A*, **101** (39), 14300-14304.

Becker, J. (2000). *The Chinese*. London: John Murray.

Beerli, P. (2004). Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol Ecol*, **13** (4), 827-836.

Beerli, P., & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152** (2), 763-773.

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A*, **98** (8), 4563-4568.

Behar, D. M., Thomas, M. G., Skorecki, K., Hammer, M. F., *et al.* (2003). Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. *Am J Hum Genet*, **73** (4), 768-779.

Bittles, A., & Chew, Y.-Y. (1998). Eugenics and population policies. In S. S. Strickland & P. S. Shetty (Eds.), *Human Biology and Social Inequality*. Cambridge: Cambridge University Press. pp 272-287.

Black, M. L. (2000). *Analysis of the Population Genetics of the Han and Hui of Liaoning Province, Peoples Republic of China*. MSc Thesis Edith Cowan University, Perth, Australia.

Black, M. L., Wang, W., & Bittles, A. H. (2001). A genome-based study of the Muslim Hui community and the Han population of Liaoning Province, PR China. *Hum Biol*, **73** (6), 801-813.

Bosch, E., Calafell, F., Perez-Lezuan, A., *et al.* (2000). Y chromosome STR haplotypes in four populations from northwest Africa. *Int J Legal Med.* **114**(1-2), 36-40.

Bosch, E., Calafell, F., Comas, D., Oefner, P. J., *et al.* (2001). High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet*, **68** (4), 1019-1029.

Bosch, E., Lee, A., Calafell, F., Arroyo, E., *et al.* (2002). High resolution Y-chromosome typing: 19 STRs amplified in three multiplex reactions. *Forensic Sci Int*, **125**, 42-51.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., *et al.* (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, **368** (6470), 455-457.

Box, J. F. (1978). *R.A. Fisher : the Life of a Scientist*. New York: Wiley.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, **234** (2), 177-186.

Campbell, C., Lee, J. E., & Elliot, M. (2002). Identity construction and reconstruction: naming and Manchu ethnicity in North East China, 1749-1909. *Hist. Methods*, **35** (3), 101-115.

Cann, R. L. (2001). Genetic clues to dispersal in human populations: retracing the past from the present. *Science*, **291** (5509), 1742-1748.

Capelli, C., Wilson, J. F., Richards, M., Stumpf, M. P., *et al.* (2001). A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am J Hum Genet*, **68** (2), 432-443.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., *et al.* (1985). A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science*, **230** (4732), 1403-1406.

Cavalli-Sforza, L. L. (1998a). The Chinese human genome diversity project. *Proc Natl Acad Sci U S A*, **95** (20), 11501-11503.

Cavalli-Sforza, L. L. (1998b). [Man and the diversity of his genome. An extraordinary phase in the history of population genetics]. *Pathol Biol (Paris)*, **46** (2), 98-102. [In French].

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton, N.J: Princeton University Press.

Cavalli-Sforza, L. L., Wilson, A. C., Cantor, C. R., Cook-Deegan, R. M., *et al.* (1991). Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics*, **11** (2), 490-491.

Cavalli-Sforza, L. L. and Bodmer, W. F. (1971). *The Genetics of Populations*. San Fransisco, W.H. Freeman.

Chaix, R., Austerlitz, F., Morar, B., Kalaydjieva, L., *et al.* (2004). Vlax Roma history: what do coalescent-based methods tell us? *Eur J Hum Genet*, **12** (4), 285-292.

Chen, Y. S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A. S., *et al.* (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet*, **57** (1), 133-149.

Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., *et al.* (1998). Genetic relationship of populations in China. *Proc Natl Acad Sci U S A*, **95** (20), 11763-11768.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution*, **23**, 72-84.

Cockerham, C. C., & Weir, B. S. (1973). Descent measures for two loci with some applications. *Theor Popul Biol*, **4** (3), 300-330.

Coop, G., & Griffiths, R. C. (2004). Ancestral inference on gene trees under selection. *Theor Popul Biol*, **66** (3), 219-232.

Cooper, G., Amos, W., Hoffinan, D., & Rubinsztein, D. C. (1996). Network analysis of human Y microsatellite haplotypes. *Hum Mol Genet*, **5** (11), 1759-1766.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., *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** (5). 1197-1214.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. London, John Murray.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex.* London, John Murray.

de Knijff, P., Kayser, M., Caglia, A., Corach, D., *et al.* (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med*, **110** (3), 134-149.

Deng, W., Shi, B., He, X., Zhang, Z., *et al.* (2004). Evolution and migration history of the Chinese population inferred from Chinese Y-chromosome evidence. *J Hum Genet.* **49** (7), 339-348.

Diamond, N. (1995). Defining the Miao: Ming, Qing and contemporary views. In S. Harrell (Ed.), *Cultural Encounters on China's Ethnic Frontiers.* Seattle: University of Washington Press.

Dikotter, F. (1992). *The Discourse of Race in Modern China.* Stanford University Press, Stanford University.

Ding, Y. C., Wooding, S., Harpending, H. C., Chi, H. C., *et al.* (2000). Population structure and history in East Asia. *Proc Natl Acad Sci U S A*, **97** (25), 14003-14006.

Donnelly, P., & Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annu Rev Genet*, **29**, 401-421.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161** (3), 1307-1320.

Du, R., & Yip, V. F. (1993). *Ethnic Groups in China.* Beijing and New York: Science Press.

Du, R. B., Zhao, Z.L., Xu, L.J., Wang, Y.F., *et al.* (1981). Percentage and types of consanguineous marriages of different nationalities and regions in China. *Nat Med J China*, **61**, 723-728 [In Chinese].

Dupuy, B. M., Stenersen, M., Egeland, T., & Olaisen, B. (2004). Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum Mutat*, **23** (2), 117-124.

Ebrey, P. B. (1999). *The Cambridge Illustrated history of China.* Cambridge : Cambridge University Press.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor Popul Biol*, **3** (1), 87-112.

Excoffier, L. (2001). Analysis of population subdivision. In D. Balding,M. Bishop, and C. Cannings. (Ed.), *Handbook of Statistical Genetics*. Chichester UK: John Wiley & Sons, pp. 271-307.

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131** (2), 479-491.

Fairbank, J & L. Reicher (1990). *China: Transformation and Transition*. Allen & Unwin Australia, Sydney.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164** (4), 1567-1587.

Festini, F. & de Martino, M. (2004). Twenty five years of the one child family policy in China. *J Epidem Commun Hlth,* **58**, 358-360.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon.

Fix, A. G. (1997). Gene frequency clines produced by kin-structured founder effects. *Hum Biol*, **69** (5), 663-673.

Fix, A. G. (2004). Kin-structured migration: causes and consequences. *Am J Hum Biol*, **16** (4), 387-394.

Forster, P., Harding, R., Torroni, A., & Bandelt, H. J. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet*, **59** (4), 935-945.

Fu, Y. X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, **147** (2), 915-925.

Gagneux, P., Boesch, C., & Woodruff, D. S. (1999). Female reproductive strategies, paternity and community structure in wild West African chimpanzees. *Anim Behav*, **57** (1), 19-32.

Gebo, D. L., Dagosto, M., Beard, K. C., Qi, T., *et al.* (2000). The oldest known anthropoid postcranial fossils and the early evolution of higher primates. *Nature*, **404** (6775), 276-278.

Gladney, D. C. (1996). *Muslim Chinese: Ethnic Nationalism in the Peoples Republic.* Cambridge, Massachusetts: Harvard University Press.

Gladney, D. C. (1998). *Ethnic Identity in China: The Making of a Muslim Minority* nationality. Fort Worth, Texas: Harcourt Brace & Company.

Gladney, D. C. (2002). Xinjiang: China's future West Bank ? *Current History*, **101**, 267-270.

Goldstein, D. B., & Chikhi, L. (2002). Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet*, **3**, 129-152.

Goldstein, D. B., & Pollock, D. D. (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J Hered*, **88** (5), 335-342.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L., & Feldman, M. W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139** (1), 463-471.

Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Linares, A. R., *et al.* (1996). Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol Biol Evol*, **13** (9), 1213-1218.

Gresham, D., Morar, B., Underhill, P. A., Passarino, G., *et al.* (2001). Origins and divergence of the Roma (gypsies). *Am J Hum Genet*, **69** (6), 1314-1331.

Griffin, T. J., & Smith, L. M. (2000). Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends Biotechnol*, **18** (2), 77-84.

Grimes, B. (2003). *The Ethnologue: Languages of the World.* Summer Institute of Linguistics. www.ethnologue.com.

Guo, S. W., & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48** (2), 361-372.

Hammer, M. F., & Horai, S. (1995). Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet*, **56** (4), 951-962.

Hammer, M. F., Karafet, T. M., Redd, A. J., Jarjanazi, H., *et al.* (2001). Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol*, **18** (7), 1189-1203.

Hammond, R. L., & Keller, L. (2004). Conflict over male parentage in social insects. *PLoS Biol*, **2** (9), E248.

Harding, R. M., & McVean, G. (2004). A structured ancestral population for the evolution of modern humans. *Curr Opin Genet Dev*, **14** (6), 667-674.

Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., *et al.* (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet*, **60** (4), 772-789.

Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., *et al.* (1998). Genetic traces of ancient demography. *Proc Natl Acad Sci U S A*, **95** (4), 1961-1967.

Hartl, D. L., & Clark, A. G. (1997). *Principles of Population Genetics* (3rd ed.). Sunderland, MA: Sinauer Associates.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Hedrick, P. W. (2000). *Genetics of Populations* (2nd ed.). Sudbury, Mass.: Jones and Bartlett Publishers.

Herrnstadt, C., Elson, J. L., Fahy, E., Preston, G., *et al.* (2002). Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet*, **70** (5), 1152-1171.

Horai, S., Kondo, R., Nakagawa-Hattori, Y., Hayashi, S., *et al.* (1993). Peopling of the Americas, founded by four major lineages of mitochondrial DNA. *Mol Biol Evol*, **10** (1), 23-47.

Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., *et al.* (1996). mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. *Am J Hum Genet*, **59** (3), 579-590.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J Ed Psych*, **24**, 417-441.

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol*, **23** (2), 183-201.

Hudson, R. R. (1990). Gene geneologies and the coalescent process. *Oxford Sur Evo Biol*, **7**, 1-44.

Iriondo, M., Barbero, M. C., & Manzano, C. (2003). DNA polymorphisms detect ancient barriers to gene flow in Basques. *Am J Phys Anthropol*, **122** (1), 73-84.

Jobling, M., Hurles, M., & Tyler-Smith, C. (2004). *Human Evolutionary Genetics: Origins, Peoples and Disease*. New York: Garland Science.

Jobling, M. A., & Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. *Trends Genet*, **11** (11), 449-456.

Jobling, M. A., & Tyler-Smith, C. (2000). New uses for new haplotypes: the human Y chromosome, disease and selection. *Trends Genet*, **16** (8), 356-362.

Karafet, T., Xu, L., Du, R., Wang, W., *et al.* (2001). Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet*, **69** (3), 615-628.

Karafet, T. M., Zegura, S. L., Posukh, O., Osipova, L., *et al.* (1999). Ancestral Asian source(s) of New World Y-chromosome founder haplotypes. *Am J Hum Genet*, **64** (3), 817-831.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., *et al.* (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med*, **110** (3), 125-133, 141-129.

Kayser, M., Brauer, S., Weiss, G., Schiefenhovel, W., *et al.* (2001a). Independent histories of human Y chromosomes from Melanesia and Australia. *Am J Hum Genet*, **68** (1), 173-190.

Kayser, M., Krawczak, M., Excoffier, L., Dieltjes, P., *et al.* (2001b). An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am J Hum Genet*, **68** (4), 990-1018.

Kimura, M. (1968). Evolutionary rate at a molecular level. *Nature*, **217**, 624-626.

King, J. L., & Jukes, T. H. (1969). Non-Darwinian evolution. *Science*, **164** (881), 788-798.

Kingman, J. F. C. (1982). On the genealogy of large populations. *J App Prob*, **19A**, 27-43.

Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., *et al.* (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet*, **72** (2), 313-332.

Kivisild, T., Tolk, H. V., Parik, J., Wang, Y., *et al.* (2002). The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol*, **19** (10), 1737-1751.

Knowles, L. L. (2004). The burgeoning field of statistical phylogeography. *J Evol Biol*, **17** (1), 1-10.

Kwak, K., Jin, H., Shin, D., *et al.* (2005) Y-chromosome STR haplotypes and their applications to forensic and population studies in east Asia. *Int J Legal Med.* **119**, 195-201.

Lell, J. T., Sukernik, R. I., Starikovskaya, Y. B., Su, B., *et al.* (2002). The dual origin and Siberian affinities of Native American Y chromosomes. *Am J Hum Genet*, **70** (1), 192-206.

Labuda, D., Zietkiewicz, E. and Yotova, V. (2000). Archaic lineages in the history of modern humans. *Genetics*. **156** (2), 799-808.

Leslie, D. D. (1986). *Islam in Traditional China: a Short History to 1800*. Canberra: Canberra College of Advanced Education.

Lipman, J. N. (1997). *Familiar Strangers: a History of Muslims in Northwest China*. Hong Kong University: Hong Kong University Press.

Litzinger, R. A. (1995). Making histories: contending conceptions of the Yao past. In S. Harrell (Ed.), *Cultural Encounters on China's Ethnic Frontiers*. Seattle: University of Washington Press.

Ma, J. Z. (1996). The role of Islam in the formation of the culture and economy of the Hui community in Guangzhou. *J Muslim Minority Affairs*, **16**(1), 31-39.

Macaulay, V., Richards, M., Hickey, E., Vega, E., *et al.* (1999). The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet*, **64** (1), 232-249.

Marjoram, P., & Donnelly, P. (1994). Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics*, **136** (2), 673-683.

Melton, T., Clifford, S., Kayser, M., Nasidze, I., *et al.* (2001). Diversity and heterogeneity in mitochondrial DNA of North American populations. *J Forensic Sci*, **46** (1), 46-52.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., *et al.* (1953). Equations of state calculations by fast computing machines. *J Chem Phys*, **21**, 1087-1091.

Michalakis, Y. and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, **142** (3), 1061-1064.

Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., *et al.* (1994). Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, **265** (5176), 1193-1201.

Moser, L. J. (1985). *The Chinese Mosiac: The Peoples and Provinces of China.* Boulder, Colorado: Westview Press.

National Bureau of Statistics of China (2002). *China Population Statistics Handbook.* Beijing: China Statistics Press.

Nebel, A., Fildon, D., Hohoff, C., *et al.*, (2001). Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and DYS392. *Eur J Hum Genet.* **9** (1), 22-26.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A*, **70** (12), 3321-3323.

Nei, M. (1977). F-statistics and analysis of gene diversity in subdivided populations. *Ann Hum Genet*, **41** (2), 225-233.

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

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics.* Oxford: Oxford University Press.

Nei, M., & Roychoudhury, A. K. (1973). Probability of fixation and mean fixation time of an overdominant mutation. *Genetics*, **74** (2), 371-380.

Nei, M., & Roychoudhury, A. K. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics*, **76** (2), 379-390.

Nei, M., & Roychoudhury, A. K. (1993). Evolutionary relationships of human populations on a global scale. *Mol Biol Evol*, **10** (5), 927-943.

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol*, **40** (1), 56-63.

Overall, A. D., Ahmad, M., & Nichols, R. A. (2002). The effect of reproductive compensation on recessive disorders within consanguineous human populations. *Heredity*, **88** (6), 474-479.

Overall, A. D., Ahmad, M., Thomas, M. G., & Nichols, R. A. (2003). An analysis of consanguinity and social structure within the UK Asian population using microsatellite data. *Ann Hum Genet*, **67** (Pt 6), 525-537.

Overall, A. D., & Nichols, R. A. (2001). A method for distinguishing consanguinity and population substructure using multilocus genotype data. *Mol Biol Evol*, **18** (11), 2048-2056.

Page, R. D. M., & Holmes, E. C. (1998). *Molecular Evolution: A Phylogenetic Approach*. London: Blackwell Science Ltd.

Paludan, A. (1998). *Chronicle of the Chinese Emperors*. London: Thames & Hudson.

Peters, H. A. (2001). Ethnicity along China's southwestern frontier. *J East Asian Arch*, **3**, 75-103.

Population Reference Bureau. (2004). *World Population Data Sheet*. Washington D.C.

Posada, D., & Crandall, K. A. (2001). Intraspecific gene genealogies: trees grafting into networks. *TREE*, **16** (1), 37-45.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*, **16** (12), 1791-1798.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155** (2), 945-959.

Qian, Y., Qian, B., Su, B., Yu, J., *et al.* (2000). Multiple origins of Tibetan Y chromosomes. *Hum Genet*, **106** (4), 453-454.

Qian, Y. P., Chu, Z. T., Dai, Q., Wei, C. D., *et al.* (2001). Mitochondrial DNA polymorphisms in Yunnan nationalities in China. *J Hum Genet*, **46** (4), 211-220.

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., *et al.* (2004). Where West meets East: the complex mtDNA landscape of the Southwest and Central Asian corridor. *Am J Hum Genet*, **74** (5), 827-845.

Quintana-Murci, L., Krausz, C., Zerjal, T., Sayar, S. H., *et al.* (2001). Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *Am J Hum Genet*, **68** (2), 537-542.

Richards, M., Macaulay, V., Hickey, E., Vega, E., *et al.* (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet*, **67** (5), 1251-1276.

Richards, M. B., Macaulay, V. A., Bandelt, H. J., & Sykes, B. C. (1998). Phylogeography of mitochondrial DNA in Western Europe. *Ann Hum Genet*, **62** (Pt 3), 241-260.

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., *et al.* (1996). Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. *Hum Mol Genet*, **5** (7), 1029-1033.

Rogers, A. R., & Harpending, H. (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol*, **9** (3), 552-569.

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet*, **3** (5), 380-390.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., *et al.* (2002). Genetic structure of human populations. *Science*, **298** (5602), 2381-2385.

Rousset, F., & Raymond, M. (1995). Testing heterozygote excess and deficiency. *Genetics*, **140** (4), 1413-1419.

Ruiz Linares, A., Nayar, K., Goldstein, D. B., Hebert, J. M., *et al.* (1996). Geographic clustering of human Y-chromosome haplotypes. *Ann Hum Genet*, **60 (Pt 5)**, 401-408.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4** (4), 406-425.

Salas, A., Richards, M., De la Fe, T., Lareu, M. V., *et al.* (2002). The making of the African mtDNA landscape. *Am J Hum Genet*, **71** (5), 1082-1111.

Satta, Y., & Takahata, N. (2002). Out of Africa with regional interbreeding? Modern human origins. *Bioessays*, **24** (10), 871-875.

Semino, O., Magri, C., Benuzzi, G., Lin, A. A., *et al.* (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet*, **74** (5), 1023-1034.

Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., *et al.* (2000). The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science*, **290** (5494), 1155-1159.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139** (1), 457-462.

Stanford Human Genome Centre. (1998). Radiation Hybrid mapping of Chromosome 13 and 15: At: http://www-shgc.stanford.edu/RH/index.html.

Steinberg, D. (2000). Genetic variation illuminates murky human history. *The Scientist*, **14** (15), 10.

Strachan, T., & Read, A. P. (2004). *Human Molecular Genetics (3rd ed)*. New York: Garland Science.

Su, B., Xiao, C., Deka, R., Seielstad, M. T., *et al.* (2000). Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. *Hum Genet*, **107** (6), 582-590.

Su, B., Xiao, J., Underhill, P., Deka, R., *et al.* (1999). Y-chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet*, **65** (6), 1718-1724.

Tajima, A., Hayami, M., Tokunaga, K., Juji, T., *et al.* (2004). Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. *J Hum Genet*, **49** (4), 187-193.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105** (2), 437-460.

Tajima, F. (1989a). The effect of change in population size on DNA polymorphism. *Genetics*, **123** (3), 597-601.

Tajima, F. (1989b). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123** (3), 585-595.

Takezaki, N., & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, **144** (1), 389-399.

Tapp, N. (2002). In defence of the archaic: a reconsideration of the 1950s ethnic classification project in China. *Asian Ethnicity*, **3**, 63-84.

Templeton, A. (1992). Human origins and analysis of mitochondrial DNA sequences. *Science*, **255** (5045), 737.

Templeton, A. (2002). Out of Africa again and again. *Nature*, **416** (6876), 45-51.

Thangaraj, K., Sridhar, V., Kivisild, T., Reddy, A. G., *et al.* (2005). Different population histories of the Mundari- and Mon-Khmer-speaking Austro-Asiatic tribes inferred from the mtDNA 9-bp deletion/insertion polymorphism in Indian populations. *Hum Genet*, **116** (6), 507-517.

Thomas, M. G., Skorecki, K., Ben-Ami, H., Parfitt, T., *et al.* (1998). Origins of Old Testament priests. *Nature*, **394** (6689), 138-140.

Torroni, A., Bandelt, H. J., Macaulay, V., Richards, M., *et al.* (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet*, **69** (4), 844-852.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., *et al.* (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics*, **144** (4), 1835-1850.

Torroni, A., Miller, J. A., Moore, L. G., Zamudio, S., *et al.* (1994). Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. *Am J Phys Anthropol*, **93** (2), 189-199.

Torroni, A., Schurr, T. G., Cabell, M. F., Brown, M. D., *et al.* (1993). Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet*, **53** (3), 563-590.

Torroni, A., Schurr, T. G., Yang, C. C., Szathmary, E. J., *et al.* (1992). Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics*, **130** (1), 153-162.

Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., *et al.* (2005). Traces of Archaic Mitochondrial Lineages Persist in Austronesian-Speaking Formosan Populations. *PLoS Biol*, **3** (8), e247.

Underhill, P.A. (2004). A synopsis of extant Y chromosome diversity in East Asia and Oceania. In L. Sagart & R. Blench & A. Sanchez-Mazas (Eds.), *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*. London: Routledge Curzon. pp. 301-319.

Underhill, P. A. (2000). The case for an African rather than Asian origin of the human Y-chromosome YAP insertion. In L. Jin, M. Seielstad & C. Xiao (Eds.), *Genetic, Linguistic and Archeological Perspectives of Human Diversity in Southeast Asia. Rec Adv Hum Genet*, Vol. **8**, pp. 43-56.

Underhill, P. A., Jin, L., Lin, A. A., Mehdi, S. Q., *et al.* (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res*, **7** (10), 996-1005.

Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., *et al.* (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet*, **65** (Pt 1), 43-62.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., *et al.* (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet*, **26** (3), 358-361.

Valsecchi, E., Hale, P., Corkeron, P., & Amos, W. (2002). Social structure in migrating humpback whales (*Megaptera novaeangliae*). *Mol Ecol*, **11** (3), 507-518.

Vigilant, L., Pennington, R., Harpending, H., Kocher, T. D., *et al.* (1989). Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci U S A*, **86** (23), 9350-9354.

Wang, W., Qian, C. and Bittles, A.H. (2001). Consanguineous marriage in PR China: a study in rural Man (Manchu) communities.: *Ann Hum Biol*, **29**(6): 685-690.

Wang, W., Wise, C., Baric, T., Black, M. L., *et al.* (2003). The origins and genetic structure of three co-resident Chinese Muslim populations: the Salar, Bo'an and Dongxiang. *Hum Genet*, **113** (3), 244-252.

Ward, R. H., Frazier, B. L., Dew-Jager, K., & Paabo, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci U S A*, **88** (19), 8720-8724.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7 (2), 256-276.

Weale, M. E., Weiss, D. A., Jager, R. F., Bradman, N., *et al.* (2002). Y chromosome evidence for Anglo-Saxon mass migration. *Mol Biol Evol*, **19** (7), 1008-1021.

Weale, M. E., Yepiskoposyan, L., Jager, R. F., Hovhannisyan, N., *et al.* (2001). Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group. *Hum Genet*, **109** (6), 659-674.

Weir, B. S. (1996). *Genetic Data Analysis : Methods for Discrete Population Genetic Data* (2nd ed.). Sunderland, Mass: Sinauer Associates.

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

Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., *et al.* (2001). The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A*, **98** (18), 10244-10249.

Wen, B., Li, H., Lu, D., Song, X., *et al.* (2004a). Genetic evidence supports demic diffusion of Han culture. *Nature*, **431** (7006), 302-305.

Wen, B., Xie, X., Gao, S., Li, H., *et al.* (2004b). Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am J Hum Genet*, **74** (5), 856-865.

Wilson, I. J., Weale, M., & Balding, D. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *J Royal Stat Soc. Series A (Statistics in Society)*, **166** (Part 2), 155-201.

Wilson, J. F., Weiss, D. A., Richards, M., Thomas, M. G., *et al.* (2001). Genetic evidence for different male and female roles during cultural transitions in the British Isles. *Proc Natl Acad Sci U S A*, **98** (9), 5078-5083.

Wise, C. A., Sullivan, S. G., Black, M. L., Erber, W. N., *et al.* (2005). Y-chromosome and mitochondrial DNA studies on the population structure of the christmas island community. *Am J Phys Anthropol* (To be published).

Wong, H. M., & Dajani, A. A. (1988). *Islamic Frontiers of China*. London: Scorpion.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, **16**, 97-159.

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

Xiaotong, F. (1991). Ethnic identification in China. *Thai-Yunnan Newsletter* (11), 11-24.

Y Chromosome Consortium. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res*, **12** (2), 339-348.

Yang, B. (2004). Horses, silver and cowries: Yunnan in global perspective. *J World His*, **15**, 281-323.

Yang, X., Zhang, L., Ding, M., and Zhang, Y.P. (1994). mtDNA polymorphism of the Mang and Kuchong peoples in China. *J Yunnan Uni*, **21**, 256 - 257 [In Chinese].

Yang, Z., Dong, Y., Gao, L., Cheng, B., *et al.* (2005). The distribution of Y chromosome haplogroups in the nationalities from Yunnan Province of China. *Ann Hum Biol*, **32** (1), 80-87.

Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T., *et al.* (2002a). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet*, **70** (3), 635-651.

Yao, Y. G., Nie, L., Harpending, H., Fu, Y., *et al.* (2002b). Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity. *Am J Phys Anthropol*, **188** (1), 63-76.

Yao, Y. G., & Zhang, Y. P. (2002). Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. *J Hum Genet*, **47** (6), 311-318.

Yao, Y. G., Watkins, W. S., & Zhang, Y. P. (2000). Evolutionary history of the mtDNA 9-bp deletion in Chinese populations and its relevance to the peopling of east and southeast Asia. *Hum Genet*, **107** (5), 504-512.

Yin, M. (Ed.). (1994). *China's Minority Nationalities*. Beijing: Foriegn Languages Press Beijing.

Zietkiewicz, E., Yotova, V., Gehl, D. et al. (2003) Haplotypes in the dystrophin DNA segment point to a mosaic origin of modern human diversity. *Am J. Hum. Genet.* **73** (5), 994-1015

Zerjal, T., Beckman, L., Beckman, G., Mikelsaar, A. V., *et al.* (2001). Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations. *Mol Biol Evol*, **18** (6), 1077-1087.

Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R., *et al.* (2002). A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am J Hum Genet*, **71** (3), 466-482.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., *et al.* (2003). The genetic legacy of the Mongols. *Am J Hum Genet*, **72** (3), 717-721.

Zhivotovsky, L. A., Underhill, P. A., Cinnioglu, C., Kayser, M., *et al.* (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet*, **74** (1), 50-61.

APPENDICES

Appendix A

Primer Sequences

Appendix A.1 Autosomal STR PCR primer sequences

Locus	Label	Primer Sequence
D13S126	FAM	(F) 5'-tcaccagtaaaatgctattgg-3'
		(R) 5'-gtgattttcaaatttgctctg-3'
D13S133	TET	(F) 5'-ggcaacatagggaaaccctagc-3'
		(R) 5'-gctaggactacaggtgcaaacc-3'
D13S192	HEX	(F) 5'-gggtaacatagcaagacccc-3'
		(R) 5'-aggtatgagccatctcgtcc-3'
D13S270	HEX	(F) 5'-agtgcctgggtatgaacgtg-3'
		(R) 5'-ctggaaatgccttggaagga-3'
D15S101	HEX	(F) 5'-gagccaagatcatgttgc-3'
		(R) 5'-tgcccactagtttgagaca-3'
D15S108	HEX	(F) 5'-attcttaacaggaagtgaggg-3'
		(R) 5'-aacatgagtttcagagggg-3'
D15S11	TET	(F) 5'-gacatgaacagaggtaaattggtgg-3'
		(R) 5'-gctctctaagatcactggatagg-3'
D15S97	FAM	(F) 5'-tctccctccaataatgtgac-3'
		(R) 5'-tgagtcaatgattgaaattactg-3'
D15S98	HEX	(F) 5'-catgtgaaactgcaaaagctg-3'
		(R) 5'-aaaagtcgcatttggtcgtt-3'
GABRB3	HEX	(F) 5'-ctcttgttcctgttgctttcaatacac-3',
		(R) 5'-cactgtgctagtagttcagctc-3'

Appendix A.2 Y-chromosome STR PCR primer sequences

Locus	Label	Primer Sequence
DYS 19	HEX	(F) 5'-ctactgagtttctgttatagt-3'
		(R) 5'-atggcatgtagtgaggaca-3'
DYS 388	FAM	(F) 5''-gtgagttagccgtttagcga-3'
		(R) 5'-cagatcgcaaccactgcg-3'
DYS 389	FAM	(F) 5'-ccaactctcatctgtattatctatg-3'
		(R) 5'-tcttatctccacccaccaga-3'
DYS 390	TET	(F) 5'-tatattttacacatttttgggcc-3'
		(R) 5'-tgacagtaaaatgaacacattgc-3'
DYS 391	TET	(F) 5'-ctattcattcaatcatacaccca-3'
		(R) 5'-gattctttgtggtgggtctg-3'
DYS 392	HEX	(F) 5'-tcattaatctagcttttaaaaacaa-3'
		(R) 5'-agacccagt tgatgcaatgt-3'
DYS 393	TET	(F) 5'-gtggtcttctacttgtgtcaatac-3'
		(R) 5'-aactcaagtccaaaa aatgagg-3'

Appendix A.3 Y-chromosome UEP primer sequences

Locus	PCR Primer Sequence
M1	(F) 5^1-caggggaagataaagaaata-3^1
	(R) 5^1-actgctaaaaggggatggat-3^1
M9	(F) 5^1-gcagcatataaaactttcagg-3^1
	(R) 5^1-aaaacctaactttgctcaagc-3^1
M45	(F) 5^1-attggcagtgaaaaattatagcta-3^1
	(R) 5^1- aatatgttcctgacaccttcc -3^1
M89	(F) 5^1-tataatcaagttaccaattactggc-3^1
	(R) 5^1-ttttgtaacattgaatggcaaa-3^1
M122	(F) 5^1-tggtaaactctacttagttgccttt-3^1
	(R) 5^1-cagcgaattagattttcttgc-3^1
M130	(F) 5^1-tatctcctcttctattgcag-3^1
	(R) 5^1-ccacaaggggaaaaaacac-3^1
M175	(F) 5^1-ttgagcaagaaaaatagtaccca-3^1
	(R) 5^1-ctccattcttaactatctcaggga-3^1

Appendix A.4 Mitochondrial HV-I PCR primer sequences

Forward: 5^1-tgtaaaacgacggccagttgatttcacggaggatggtg-3^1

Reverse: 5^1-caggaaacagctatgaccctccaccattagcataccgcca-3^1

Appendix A.5 Primers for mtDNA 9bp indel detection in theCOII/tRNA^{LYS} intergenic region

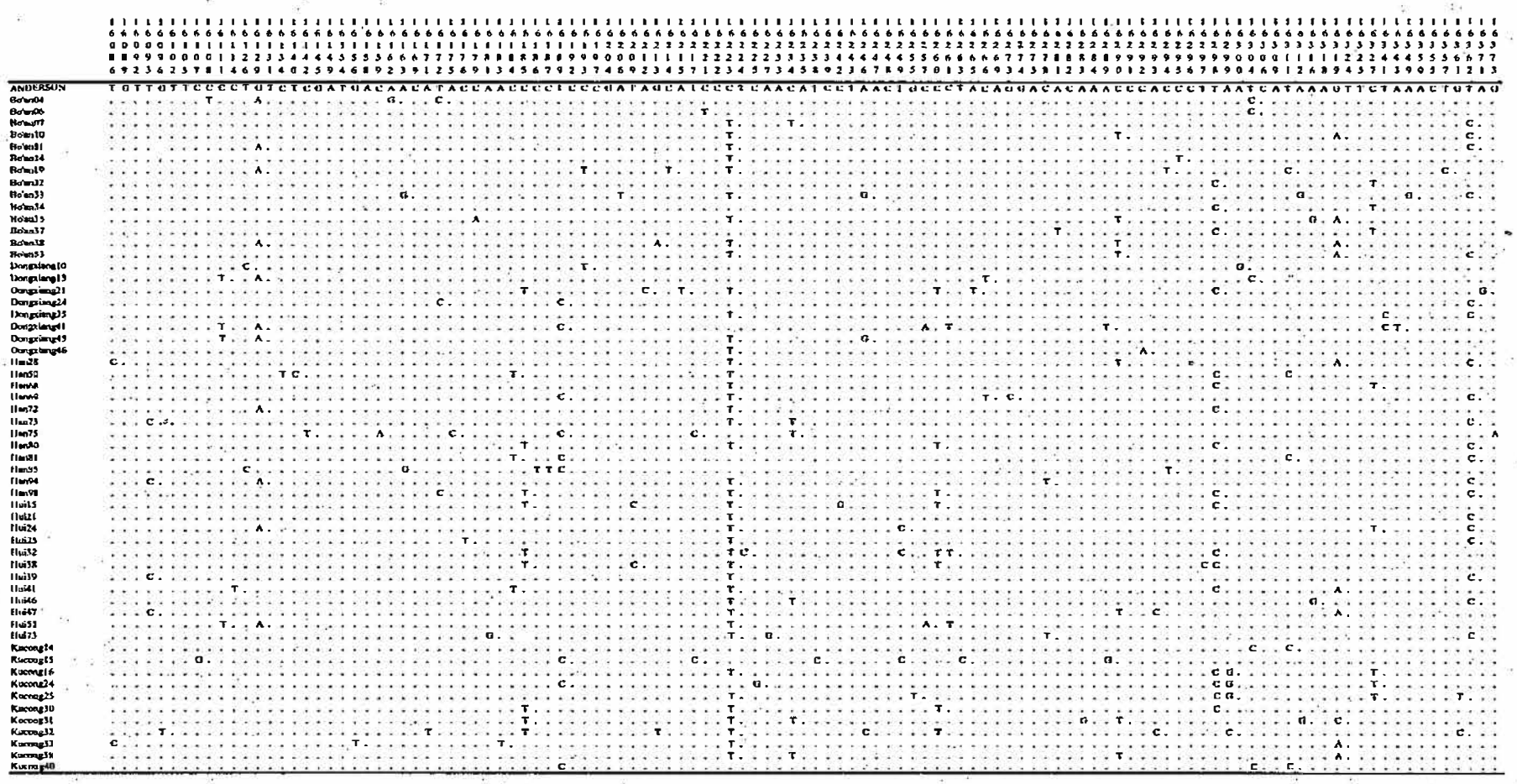
Forward 5^1 - tcgtcctagaattaattccc – 3^1

Reverse 5^1 - agttagctttacagtgggct-3^1

Appendix B

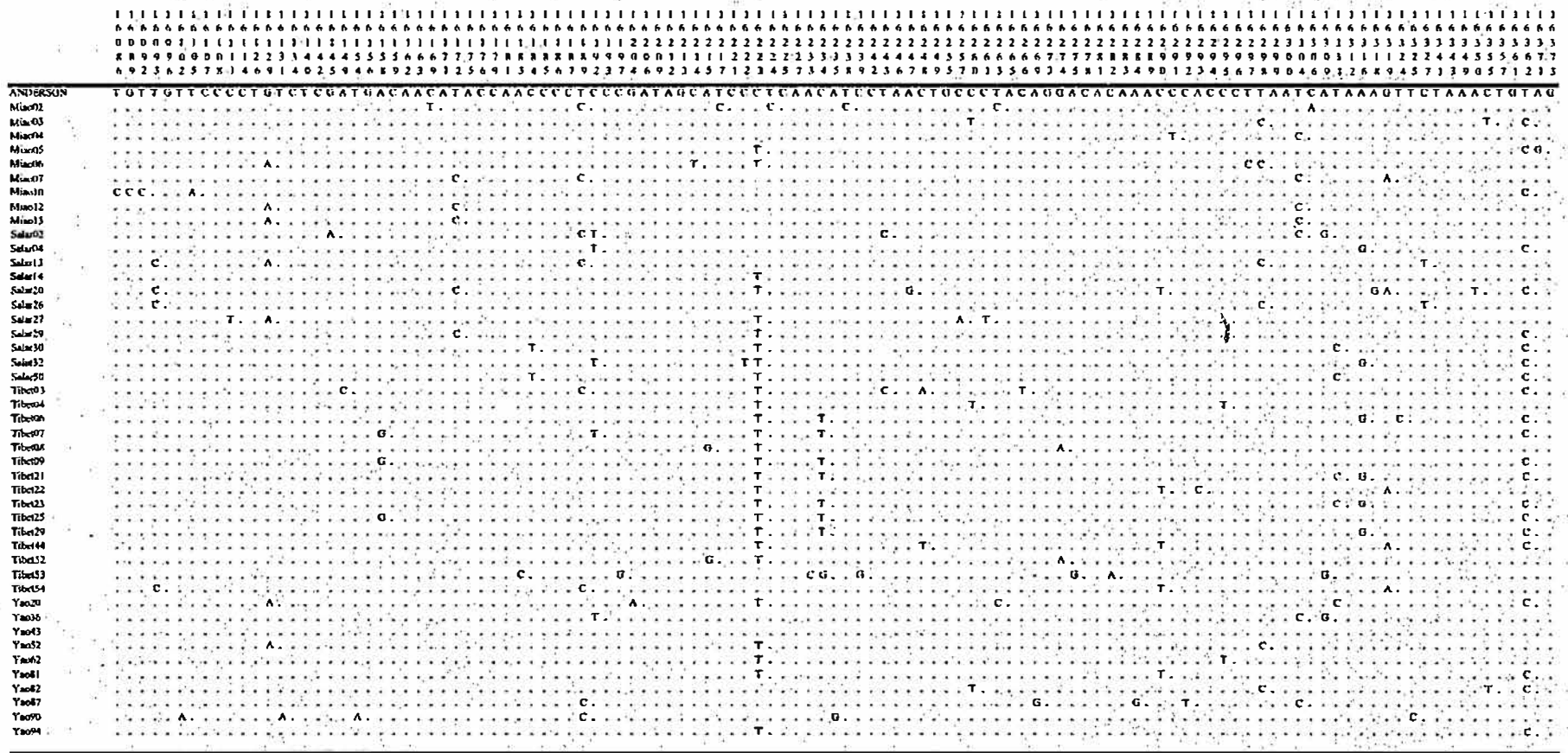
Relevant raw data derived from the genomic analyses

Appendix B.1 mtDNA HV1 sequences from the 9 study populations



233

Appendix B.1 (continued)



* Sequences compared with Anderson sequence as control (Anderson et al., 1981). Mutation from control notated as relevant base. Only positions at which mutations were found in a least one individual are shown.

234

Appendix B.2 Y-chromosome STR haplotypes

B.2.1 Bo'an Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Bo'an27	14	10	12	16	12
Bo'an22	14	12	12	16	12
Bo'an28	14	12	12	16	12
Bo'an05	14	12	12	17	12
Bo'an29	14	12	12	17	12
Bo'an38	14	12	12	18	12
Bo'an10	14	12	12	19	13
Bo'an34	14	12	12	17	14
Bo'an20	14	12	12	18	14
Bo'an02	14	13	12	15	12
Bo'an07	14	12	13	16	12
Bo'an12	14	12	13	16	12
Bo'an04	14	12	13	17	12
Bo'an01	14	12	13	16	13
Bo'an19	14	12	13	17	13
Bo'an32	14	12	13	19	13
Bo'an40	14	12	13	16	15
Bo'an18	14	13	13	19	13
Bo'an37	14	14	13	16	12
Bo'an31	14	14	13	17	12
Bo'an30	14	14	13	17	13
Bo'an15	17	12	12	17	12
Bo'an06	17	12	13	15	12
Bo'an11	17	12	13	15	12
Bo'an36	17	12	13	17	12
Bo'an03	17	12	13	17	13
Bo'an23	17	12	13	17	13
Bo'an21	17	12	13	19	13
Bo'an25	17	12	13	15	14
Bo'an14	17	12	13	17	14
Bo'an39	17	12	13	17	14
Bo'an09	17	13	13	15	13
Bo'an16	17	13	13	18	13
Bo'an17	17	13	13	18	13
Bo'an35	17	14	13	17	12
Bo'an13	17	14	13	17	13
Bo'an24	17	14	14	17	14
Bo'an08	18	12	13	15	12
Bo'an33	18	12	13	15	12
Bo'an26	18	12	13	17	13

B.2.2 Dongxiang Y chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Dongxiang40	14	10	12	17	12
Dongxiang08	14	10	13	16	12
Dongxiang32	14	10	14	17	12
Dongxiang38	14	12	12	15	14
Dongxiang18	14	12	12	16	11
Dongxiang45	14	12	12	16	12
Dongxiang17	14	12	12	16	13
Dongxiang31	14	12	12	16	13
Dongxiang43	14	12	12	16	14
Dongxiang12	14	12	12	17	13
Dongxiang21	14	12	12	17	13
Dongxiang23	14	12	12	17	13
Dongxiang09	14	12	12	17	14
Dongxiang34	14	12	12	18	13
Dongxiang04	14	12	13	16	11
Dongxiang05	14	12	13	16	11
Dongxiang16	14	12	13	17	13
Dongxiang15	14	12	13	18	11
Dongxiang11	14	12	13	18	13
Dongxiang22	14	12	14	17	12
Dongxiang26	14	12	14	17	13
Dongxiang29	14	12	15	18	13
Dongxiang06	14	13	13	16	13
Dongxiang37	14	13	13	16	13
Dongxiang24	14	13	13	18	12
Dongxiang20	14	13	13	19	13
Dongxiang33	14	13	14	16	13
Dongxiang28	14	13	14	16	14
Dongxiang07	14	14	13	16	12
Dongxiang39	14	15	12	16	12
Dongxiang19	14	15	13	18	12
Dongxiang36	14	15	14	17	12
Dongxiang42	17	11	12	18	12
Dongxiang14	17	12	12	16	13
Dongxiang41	17	12	13	15	12
Dongxiang02	17	12	13	17	13
Dongxiang03	17	12	13	18	13
Dongxiang27	17	12	14	17	13
Dongxiang30	17	12	14	17	13
Dongxiang35	17	13	12	16	12
Dongxiang13	17	13	14	16	12

B.2.3 Han Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Han85	13	12	12	16	12
Han38	13	12	13	17	14
Han37	13	15	14	16	14
Han01	14	10	12	16	12
Han42	14	10	12	16	12
Han89	14	10	12	16	12
Han19	14	10	13	16	12
Han14	14	10	13	18	12
Han56	14	10	14	16	14
Han13	14	11	13	15	14
Han59	14	12	11	20	12
Han79	14	12	12	15	12
Han33	14	12	12	16	12
Han54	14	12	12	16	12
Han90	14	12	12	16	12
Han43	14	12	13	16	12
Han55	14	12	13	16	13
Han88	14	12	13	16	13
Han20	14	12	14	17	12
Han26	14	12	14	17	13
Han34	14	13	12	16	12
Han40	14	13	12	16	12
Han47	14	13	12	16	14
Han60	14	13	13	17	12
Han02	14	13	13	16	14
Han46	14	14	12	16	12
Han70	14	14	12	16	12
Han41	14	14	13	15	13
Han100	15	10	12	16	12
Han49	15	10	12	16	12
Han05	15	10	12	17	12
Han18	15	10	12	18	12
Han15	15	10	13	16	12
Han27	15	10	13	17	12
Han80	15	10	14	14	12
Han82	15	10	14	17	12
Han44	15	10	14	16	15
Han07	15	12	12	16	12
Han09	15	12	12	16	12
Han30	15	12	12	16	12
Han69	15	12	12	16	12
Han06	15	12	12	19	12
Han08	15	12	12	15	13
Han53	15	12	12	15	13
Han17	15	12	12	16	13

B.2.3 (continued)

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Han16	15	12	12	17	13
Han84	15	12	12	17	13
Han25	15	12	12	16	14
Han11	15	12	12	17	14
Han97	15	12	12	17	14
Han36	15	12	13	16	12
Han99	15	12	13	16	12
Han28	15	12	13	17	12
Han45	15	12	14	16	14
Han50	15	12	14	18	14
Han51	15	12	14	18	15
Han73	15	12	15	14	12
Han10	15	13	12	15	12
Han92	15	13	12	16	13
Han98	15	13	13	16	12
Han29	15	13	13	17	12
Han81	15	13	13	16	13
Han102	15	13	14	16	13
Han22	15	13	14	16	13
Han96	15	13	14	16	13
Han83	15	13	14	17	13
Han04	15	14	12	17	15
Han52	15	15	14	18	15
Han31	16	12	12	15	12
Han63	16	12	12	16	12
Han74	16	12	12	16	12
Han86	16	12	12	16	12
Han67	16	12	12	17	12
Han87	16	12	12	18	12
Han03	16	12	12	19	13
Han12	16	12	12	17	14
Han65	16	12	13	15	12
Han68	16	12	13	16	13
Han76	16	12	13	18	13
Han32	16	12	13	14	14
Han66	16	12	14	18	13
Han95	16	12	14	19	13
Han77	16	14	13	15	12
Han75	16	17	12	16	12
Han58	16	17	12	17	12
Han39	16	17	14	16	12
Han48	17	10	12	16	12
Han91	17	11	13	15	12
Han23	17	12	12	16	13
Han24	17	12	12	16	14
Han21	17	13	12	16	12

B.2.4 Hui Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Hui21	14	12	13	18	13
Hui38	14	12	13	18	13
Hui41	14	13	14	17	12
Hui45	14	15	13	17	13
Hui75	14	15	12	17	15
Hui32	14	17	13	17	12
Hui05	15	12	12	16	12
Hui73	15	12	12	16	13
Hui10	15	12	13	18	13
Hui24	15	12	13	18	13
Hui57	15	12	14	16	14
Hui76	15	13	13	16	12
Hui53	15	14	14	16	14
Hui39	15	14	13	16	15
Hui42	15	14	13	16	15
Hui23	15	16	13	16	12
Hui03	16	12	12	16	12
Hui25	16	12	13	16	13
Hui47	16	12	14	16	13
Hui77	16	12	14	17	13
Hui12	16	12	13	18	13
Hui51	17	14	12	17	12

B.2.5 Kucong Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Kucong37	12	14	13	15	13
Kucong08	13	11	12	16	12
Kucong18	13	11	14	15	13
Kucong16	13	11	14	16	13
Kucong15	13	11	14	16	14
Kucong25	13	12	13	15	13
Kucong23	13	12	14	17	12
Kucong31	13	13	13	17	13
Kucong24	13	13	13	17	14
Kucong35	13	13	13	18	14
Kucong12	13	13	14	14	14
Kucong32	13	13	14	14	14
Kucong40	13	13	14	14	14
Kucong19	14	12	12	18	14
Kucong36	14	12	13	15	13
Kucong09	14	12	13	17	14
Kucong21	14	12	14	16	14
Kucong33	14	13	13	17	14
Kucong38	14	13	14	15	14
Kucong30	14	14	12	18	13
Kucong13	16	12	14	16	14

B.2.7 Sala Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Sala22	14	10	12	19	11
Sala08	14	11	14	17	11
Sala31	14	12	14	14	10
Sala18	14	12	14	15	10
Sala44	14	12	13	16	10
Sala02	14	12	14	16	10
Sala01	14	12	14	17	10
Sala27	14	12	13	18	10
Sala28	14	12	13	18	10
Sala25	14	12	13	16	11
Sala03	14	12	14	16	11
Sala42	14	12	13	17	11
Sala06	14	12	12	19	11
Sala10	14	12	12	16	12
Sala35	14	12	13	15	13
Sala29	14	12	13	16	13
Sala11	14	12	13	17	14
Sala36	14	12	13	17	14
Sala26	14	14	13	17	11
Sala39	14	14	14	17	11
Sala37	14	14	14	16	13
Sala21	14	15	13	15	11
Sala05	14	15	13	17	11
Sala41	14	15	13	17	11
Sala14	17	12	12	16	11
Sala07	17	12	13	17	11
Sala09	17	12	12	18	11
Sala12	17	12	13	18	11
Sala17	17	12	12	16	12
Sala23	17	12	13	17	13
Sala04	17	13	13	18	11
Sala40	17	14	13	16	10

B.2.8 Tibetan Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Tibetan10	13	14	13	17	13
Tibetan13	14	12	13	17	13
Tibetan06	14	14	13	17	12
Tibetan17	15	11	14	16	13
Tibetan15	15	12	12	17	12
Tibetan25	15	12	12	18	13
Tibetan29	15	12	13	16	13
Tibetan28	15	12	14	15	13
Tibetan31	15	12	14	15	13
Tibetan40	15	12	14	15	13
Tibetan39	15	12	14	16	12
Tibetan52	15	12	14	16	12
Tibetan18	15	12	14	16	13
Tibetan19	15	12	14	16	13
Tibetan34	15	12	14	16	13
Tibetan26	15	14	13	18	13
Tibetan16	16	11	12	17	13
Tibetan51	16	11	14	17	13
Tibetan32	16	12	12	18	12
Tibetan54	16	12	14	15	13

B.2.9 Yao Y-chromosome STR haplotypes

	DYS19	DYS388	DYS389a	DYS389b	DYS393
Yao94	15	11	13	15	12
Yao36	15	14	12	16	13
Yao20	15	14	12	17	13
Yao52	15	14	12	17	13
Yao43	15	14	12	18	13
Yao84	15	14	13	16	13
Yao90	15	14	13	16	13
Yao85	15	14	13	17	13
Yao87	15	14	13	17	13
Yao83	16	12	13	17	15
Yao63	16	13	12	16	12
Yao79	18	12	11	17	12
Yao81	18	12	12	16	12

Appendix B.3 Autosomal Microsatellite Genotypes

B.3.1 Bo'an autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Bo'an1	105	167	113	82	259	170	153	91	144	185
	107	173	121	82	263	184	155	81	144	185
Bo'an2	101	129	103	82	243	172	147	107	144	185
	109	171	109	90	251	186	157	113	144	195
Bo'an3	105	133	103	82	243	174	147	107	144	185
	105	175	117	82	259	180	157	89	144	185
Bo'an4	109	131	97	90	243	178	143	107	144	185
	111	173	97	92	243	184	157	119	156	185
Bo'an5	105	135	97	82	249	182	157	105	144	185
	105	157	105	92	261	184	157	109	154	185
Bo'an6	105	135	103	82	243	182	147	109	144	187
	107	177	107	82	243	186	157	127	144	187
Bo'an7	105	133	97	92	243	182	149	105	144	183
	109	161	99	92	243	184	157	109	144	183
Bo'an8	103	131	97	82	243	176	157	107	152	181
	107	175	113	92	243	176	157	107	152	183
Bo'an9	101	141	97	82	245	178	155	107	144	185
	107	141	111	90	245	186	157	107	144	201
Bo'an10	101	135	103	82	243	174	147	113	130	183
	103	135	109	92	263	180	157	113	146	183
Bo'an11	103	173	109	82	243	180	145	113	144	187
	109	175	121	90	263	180	153	113	144	187
Bo'an12	103	131	97	78	243	182	145	107	144	181
	107	133	119	92	249	182	153	109	144	181
Bo'an13	107	131	103	82	243	186	147	107	144	185
	107	131	105	82	243	188	147	111	156	185
Bo'an14	103	129	103	78	241	178	155	109	144	183
	107	133	107	92	243	180	169	113	156	187
Bo'an15	103	131	97	82	243	182	147	101	142	183
	107	171	119	90	249	184	157	113	160	185
Bo'an16	103	175	109	82	243	178	147	101	142	183
	105	175	109	90	243	182	157	113	154	183
Bo'an17	107	131	103	82	249	184	143	107	144	181
	107	131	111	82	257	184	153	113	146	185
Bo'an18	105	131	107	82	243	182	153	109	142	183
	107	131	115	90	243	182	153	109	142	183
Bo'an19	103	131	107	76	243	180	155	107	142	183
	103	133	107	90	243	184	155	113	154	185
Bo'an20	109	131	117	82	243	182	147	111	138	185
	109	131	117	90	243	182	157	115	142	185

B.3.1 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Bo'an21	103	173	103	82	243	172	143	101	142	185
	109	173	103	90	243	186	147	101	142	185
Bo'an22	103	125	97	90	243	182	147	107	142	183
	109	131	109	90	259	186	157	113	142	195
Bo'an23	105	177	97	82	243	176	147	107	156	183
	111	179	103	90	249	184	157	113	156	183
Bo'an24	103	171	105	82	243	176	147	109	142	183
	103	173	105	90	243	176	147	111	156	193
Bo'an25	107	133	113	70	243	176	147	107	142	181
	109	133	115	82	261	182	147	109	142	183
Bo'an26	101	131	103	82	243	182	157	109	142	183
	103	131	113	90	265	182	167	111	142	185
Bo'an27	105	125	97	90	243	182	147	101	142	191
	107	131	115	82	259	186	157	107	142	195
Bo'an28	103	179	113	82	243	182	145	109	158	179
	103	181	117	82	243	186	147	109	162	193
Bo'an29	101	131	107	82	243	180	151	105	142	183
	101	131	119	90	259	180	157	109	160	187
Bo'an30	107	135	99	82	243	170	147	107	144	183
	107	185	117	92	259	178	155	107	156	183
Bo'an31	101	131	105	82	243	176	153	101	142	177
	103	131	117	92	269	180	161	109	144	183
Bo'an32	103	177	97	82	243	182	147	105	136	183
	103	179	115	90	243	186	157	109	142	183
Bo'an33	103	131	105	80	243	174	159	107	142	183
	105	177	109	80	255	184	161	113	142	183
Bo'an34	103	135	97	80	243	182	147	107	142	181
	105	135	113	80	249	184	153	107	142	183
Bo'an35	103	173	103	82	241	178	153	101	142	183
	109	177	111	82	243	184	157	107	146	183
Bo'an36	103	175	109	82	243	180	147	105	144	183
	103	175	111	92	243	182	155	107	146	183
Bo'an37	103	131	107	90	243	174	153	109	156	185
	107	131	109	96	249	180	153	109	158	185
Bo'an38	103	175	97	82	257	180	153	107	158	183
	109	181	103	82	265	184	171	113	158	185
Bo'an39	103	133	103	70	243	174	147	101	144	183
	109	133	113	90	243	182	157	113	156	183
Bo'an40	103	131	97	82	243	182	159	99	154	183
	107	131	117	90	261	186	163	99	158	185
Bo'an41	103	131	111	88	241	172	147	105	142	183
	107	179	117	92	243	182	153	109	142	183
Bo'an42	103	135	99	82	241	180	147	109	142	183
	107	177	105	92	243	182	153	113	156	183
Bo'an43	103	133	99	90	243	176	147	107	146	183
	103	175	119	96	259	186	157	115	156	187

B.3.1 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Bo'an44	101	167	97	82	243	176	157	107	144	185
	107	167	99	90	261	182	157	107	144	185
Bo'an45	103	131	117	90	243	176	145	99	146	185
	109	131	119	90	259	184	147	105	158	185
Bo'an46	101	131	113	82	243	176	153	107	154	183
	107	131	113	82	251	180	177	107	160	193
Bo'an47	105	171	103	90	243	184	147	107	142	173
	105	183	103	90	249	184	173	111	156	185
Bo'an48	107	179	105	82	243	174	147	109	144	181
	107	183	109	90	261	176	161	111	150	183
Bo'an49	103	165	109	82	243	178	153	99	150	183
	103	165	109	90	243	180	153	109	158	183
Bo'an50	107	173	103	82	241	176	153	101	142	181
	111	177	107	82	241	182	153	105	142	181
Bo'an51	103	131	97	82	241	180	163	99	144	185
	103	175	99	82	249	184	163	109	144	185
Bo'an52	103	131	103	82	243	178	?	105	142	185
	107	133	103	92	243	182	?	115	156	199
Bo'an53	103	131	103	82	243	176	153	99	142	183
	105	131	113	92	249	180	167	105	142	197
Bo'an54	103	131	103	82	243	182	157	107	144	183
	107	155	105	90	249	182	171	115	146	183
Bo'an55	105	131	103	82	243	170	171	107	142	185
	107	131	117	90	249	178	171	107	154	185
Bo'an56	105	171	103	82	243	184	147	107	142	183
	107	187	111	82	243	186	163	107	156	183
Bo'an57	107	131	103	82	243	178	147	105	142	183
	111	175	109	82	243	184	163	107	156	191
Bo'an58	101	163	103	82	241	182	159	101	142	183
	107	175	115	90	243	184	159	107	142	183
Bo'an59	103	133	99	82	243	178	157	101	138	179
	103	175	109	92	249	182	159	113	150	185
Bo'an60	101	131	97	82	243	174	157	101	144	181
	107	175	105	82	259	186	159	101	144	183
Bo'an61	105	131	97	82	243	180	159	107	144	183
	107	177	107	82	243	184	159	107	152	183
Bo'an62	103	131	105	90	243	178	145	107	142	183
	103	131	109	92	249	186	157	107	142	193
Bo'an63	103	131	103	82	243	178	157	105	142	183
	103	133	103	92	249	182	157	115	142	185
Bo'an64	103	133	113	82	243	180	157	107	142	183
	107	133	115	92	259	184	159	107	142	195
Bo'an65	103	131	97	82	243	180	147	107	144	179
	107	177	115	82	243	184	157	107	158	201
Bo'an66	107	131	103	82	249	176	?	107	144	179
	109	177	111	82	249	182	?	109	144	183
Bo'an67	103	173	101	82	243	174	157	109	158	183
	107	173	115	82	249	176	159	115	158	183

B.3.2 Dongxiang autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Dongxiang1	109	123	99	86	243	174	145	107	144	183
	109	123	113	86	243	176	157	107	144	185
Dongxiang2	105	117	107	80	243	170	147	107	158	181
	109	125	113	90	243	182	153	109	158	185
Dongxiang3	105	123	107	74	243	178	157	107	144	183
	109	131	109	74	243	184	171	113	144	187
Dongxiang4	105	129	97	90	241	180	153	101	144	185
	105	131	99	90	243	182	159	107	144	185
Dongxiang5	105	131	107	82	241	176	157	107	144	185
	107	133	109	92	243	182	157	113	156	185
Dongxiang6	109	131	97	90	243	182	157	101	138	181
	109	133	109	92	259	182	157	107	144	185
Dongxiang7	105	131	99	74	243	182	147	101	146	183
	105	133	105	74	263	184	157	107	160	185
Dongxiang8	109	171	103	82	243	174	147	101	136	185
	111	173	111	82	243	178	147	107	136	195
Dongxiang9	109	131	97	82	259	184	155	107	144	185
	109	131	99	82	259	184	173	113	156	185
Dongxiang10	109	171	97	78	243	182	157	107	138	185
	111	175	113	78	243	186	157	113	144	189
Dongxiang11	105	123	97	82	243	182	153	107	146	185
	111	131	103	90	259	184	157	113	146	187
Dongxiang12	105	131	103	82	243	182	147	101	144	185
	111	131	103	88	243	182	153	107	158	187
Dongxiang13	107	131	103	82	243	182	147	107	144	181
	111	131	115	82	243	184	161	111	144	185
Dongxiang14	107	131	97	90	243	184	147	107	138	181
	109	131	97	90	243	186	157	113	144	193
Dongxiang15	105	171	103	82	241	182	159	101	144	187
	111	173	107	82	243	182	159	107	144	187
Dongxiang16	105	131	99	90	243	182	159	99	138	185
	109	131	113	92	243	184	167	107	138	185
Dongxiang17	107	125	97	90	243	182	157	101	144	185
	111	131	113	90	243	182	157	107	158	187
Dongxiang18	105	125	103	82	241	182	149	99	144	187
	109	125	107	90	243	182	159	107	144	187
Dongxiang19	109	127	109	82	243	174	147	107	144	181
	109	131	109	90	249	182	147	111	156	185
Dongxiang20	109	131	103	82	257	182	153	107	144	185
	109	133	113	90	263	182	157	109	146	187

B.3.2 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Dongxiang21	105	131	97	82	243	172	147	107	156	185
	105	133	103	82	249	178	151	115	158	185
Dongxiang22	103	171	103	82	243	176	151	99	156	185
	105	173	103	90	259	186	153	107	158	195
Dongxiang23	105	169	97	82	243	182	153	101	146	185
	107	171	105	90	243	182	159	107	158	187
Dongxiang24	107	125	97	82	243	178	147	107	144	185
	111	127	103	84	259	182	157	115	150	187
Dongxiang25	103	131	99	80	243	176	153	107	144	183
	105	131	107	82	249	180	153	107	144	185
Dongxiang26	105	123	107	82	243	176	151	107	144	181
	109	131	115	90	243	182	153	107	144	187
Dongxiang27	103	123	103	82	249	180	151	107	160	185
	111	131	109	90	259	182	163	113	162	185
Dongxiang28	105	123	103	82	249	182	153	101	144	185
	113	131	111	92	259	184	157	107	158	185
Dongxiang29	109	123	103	82	243	172	147	101	138	185
	109	131	107	90	243	180	153	107	144	199
Dongxiang30	109	131	99	82	243	180	153	107	146	183
	109	131	115	90	243	180	175	109	158	185
Dongxiang31	107	131	97	82	243	186	147	101	144	185
	107	133	103	90	243	186	159	107	158	191
Dongxiang32	103	131	97	82	249	180	149	107	146	185
	109	131	103	90	249	182	153	109	158	185
Dongxiang33	105	133	97	82	243	182	157	103	144	185
	105	135	107	90	243	186	157	107	144	193
Dongxiang34	105	129	111	82	243	174	147	101	144	183
	105	131	117	90	243	180	157	107	144	185
Dongxiang35	103	131	107	82	243	178	147	99	144	185
	105	133	115	92	243	182	153	105	144	197
Dongxiang36	103	131	97	82	243	180	157	107	144	181
	105	133	103	82	259	180	157	111	158	185
Dongxiang37	107	131	103	82	243	172	149	107	164	185
	111	133	107	90	249	182	159	109	164	185
Dongxiang38	107	129	107	82	249	182	153	101	144	185
	109	131	113	90	259	184	157	107	158	185
Dongxiang39	109	127	99	82	243	176	157	101	138	181
	109	131	107	90	249	184	157	107	158	185
Dongxiang40	105	123	97	82	239	178	151	107	156	181
	111	125	117	90	243	184	175	111	158	187
Dongxiang41	107	127	103	82	243	182	153	101	144	185
	111	127	103	82	243	186	153	107	144	187
Dongxiang42	107	129	99	82	243	174	147	105	146	185
	109	131	103	90	259	186	147	107	146	185
Dongxiang43	105	129	103	82	247	172	147	101	146	187
	105	131	109	92	249	180	157	107	146	187
Dongxiang44	109	129	97	82	243	174	153	101	152	185
	109	131	105	82	245	182	157	107	156	187

B.3.2 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Dongxiang45	107	127	105	82	243	174	145	99	144	185
	109	133	111	90	243	182	159	107	158	185
Dongxiang46	105	131	97	82	241	178	149	107	148	185
	105	131	113	90	243	178	157	107	156	187
Dongxiang47	105	?	103	82	243	178	153	101	152	185
	107	?	113	90	259	182	157	107	156	197
Dongxiang55	109	127	99	82	243	182	147	123	144	?
	111	133	109	90	243	186	149	123	146	?
Dongxiang56	107	129	103	82	243	?	147	119	?	183
	111	131	103	92	243	?	155	125	?	185
Dongxiang57	107	131	105	82	243	182	153	?	?	183
	111	133	111	82	243	184	157	?	?	185
Dongxiang58	107	133	101	82	243	176	157	107	144	181
	109	135	115	92	247	178	169	113	158	187
Dongxiang59	105	?	107	82	243	178	147	123	?	185
	109	?	109	82	265	184	147	123	?	197
Dongxiang60	105	125	107	82	243	180	155	99	152	183
	109	127	113	90	249	182	159	107	156	185
Dongxiang61	103	129	99	88	243	176	153	?	?	185
	105	131	109	90	243	176	171	?	?	187
Dongxiang62	105	?	99	82	243	178	157	?	?	183
	109	?	107	90	243	172	169	?	?	185

B.3.3 Han autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Han1	111	133	103	90	243	182	147	107	142	185
	111	179	113	96	243	184	155	109	146	187
Han2	107	175	107	80	243	182	155	107	146	187
	111	179	109	90	251	184	155	107	160	187
Han3	105	135	97	80	245	178	153	101	146	187
	113	177	99	80	249	188	157	107	146	191
Han4	111	177	99	80	243	182	153	107	146	183
	111	177	103	80	243	184	153	107	160	183
Han5	111	137	103	80	243	178	147	99	146	185
	111	179	103	80	259	182	157	107	146	185
Han6	105	135	103	80	243	176	153	99	146	183
	113	137	107	80	243	186	157	107	146	187
Han7	107	179	107	80	243	172	155	107	146	187
	109	181	109	80	251	172	157	107	146	197
Han8	111	133	97	74	243	186	147	105	144	?
	113	133	97	80	243	196	153	111	146	?
Han9	107	135	107	80	243	178	153	109	144	183
	107	135	113	80	255	188	157	109	146	187
Han10	103	137	107	80	243	182	157	107	146	185
	107	139	113	92	259	182	159	107	160	185
Han11	107	?	103	80	243	182	147	105	142	187
	109	?	107	80	243	184	149	109	160	187
Han12	107	135	97	80	243	176	155	101	146	183
	111	179	107	80	243	182	157	107	158	187
Han13	107	135	97	80	243	176	147	107	146	185
	111	135	97	90	249	178	157	107	158	185
Han14	107	135	97	80	243	182	147	105	146	187
	111	179	97	80	259	182	153	109	146	187
Han15	105	135	97	80	243	182	147	105	146	187
	107	187	97	80	259	182	153	109	146	187
Han16	105	137	97	80	243	176	153	103	146	185
	105	181	97	80	243	182	157	105	160	193
Han17	113	135	87	92	243	176	151	101	146	181
	113	137	107	92	259	182	157	111	160	187
Han18	107	135	97	82	243	182	153	103	134	187
	111	135	97	90	243	182	157	113	146	187
Han19	113	135	103	?	243	176	151	101	132	?
	113	137	117	?	243	182	157	111	134	?
Han20	107	135	103	80	249	182	147	107	142	187
	107	135	117	90	259	184	157	109	146	187

B.3.3 cont

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Han21	113	135	113	?	?	176	151	?	146	?
	113	137	117	?	?	182	157	?	146	?
Han22	107	177	97	80	249	184	155	109	146	185
	107	187	97	90	259	184	157	113	146	185
Han23	109	179	97	74	243	176	153	107	146	?
	109	179	97	80	255	184	159	107	148	?
Han24	107	135	97	80	243	182	157	107	146	185
	117	181	103	80	243	182	171	107	148	191
Han25	109	179	97	?	?	184	?	107	146	?
	109	179	103	?	?	184	?	109	160	?
Han26	107	173	97	80	259	178	149	103	146	185
	109	173	105	80	259	184	157	105	146	185
Han27	107	175	103	80	243	178	147	103	146	185
	113	179	121	80	243	178	157	105	160	185
Han28	107	135	103	80	259	178	?	107	146	187
	111	183	109	80	259	182	?	107	146	191
Han29	107	135	97	80	243	180	149	107	146	187
	111	135	103	80	243	184	153	109	146	187
Han30	107	135	97	80	243	178	147	107	146	185
	113	137	103	80	243	182	157	113	166	185
Han32	?	135	103	?	243	?	?	107	146	187
	?	177	107	?	259	?	?	107	146	189
Han33	107	135	97	88	243	182	147	97	146	185
	111	187	103	92	243	182	159	109	146	185
Han34	107	135	97	80	243	184	147	105	146	185
	107	183	115	80	257	186	155	111	146	185
Han35	107	175	97	80	249	178	153	107	142	187
	111	177	105	80	249	178	157	111	146	187
Han36	107	175	?	80	?	176	147	107	142	?
	111	177	?	80	?	178	153	111	146	?
Han37	?	135	?	80	243	178	155	?	146	?
	?	137	?	80	255	178	157	?	146	?
Han38	?	135	103	80	243	176	147	107	146	187
	?	181	115	90	245	172	153	121	146	189
Han39	111	135	103	80	243	176	147	101	146	187
	111	187	103	90	249	182	157	101	146	187
Han40	111	137	103	90	243	178	147	99	144	187
	111	177	107	90	249	184	157	107	146	187
Han41	107	137	107	80	243	182	147	101	144	187
	111	181	113	90	243	182	147	107	160	187
Han42	?	135	97	80	?	174	153	?	144	?
	?	181	107	88	?	178	157	?	146	?
Han44	109	137	97	80	243	178	155	99	146	187
	111	139	97	80	243	180	157	111	146	187

B.3.3 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Han45	107	135	97	80	243	176	153	105	146	?
	111	175	115	88	249	182	157	111	146	?
Han46	107	175	97	80	243	182	147	109	146	189
	111	175	105	80	259	182	157	113	146	195
Han47	107	177	97	80	243	176	153	107	146	185
	109	177	105	80	249	182	155	113	158	191
Han48	105	135	97	80	243	182	151	107	146	183
	107	135	107	80	243	188	153	107	160	187
Han49	107	135	97	80	?	178	147	107	146	183
	113	135	109	88	?	182	155	107	150	187
Han50	107	131	97	80	249	172	147	107	146	185
	111	135	111	80	259	176	157	111	160	185
Han51	109	133	103	80	249	182	147	?	146	185
	111	179	103	80	259	182	147	?	158	185
Han52	111	179	97	80	?	182	147	105	158	185
	111	183	111	80	?	182	147	107	158	185
Han53	107	133	97	80	243	178	149	105	158	185
	111	185	111	80	243	182	157	107	158	185
Han54	111	133	97	82	243	186	147	107	146	185
	111	133	107	88	243	182	147	109	146	195
Han55	107	133	107	82	243	184	147	105	146	185
	107	185	109	82	259	184	157	107	146	195
Han56	109	133	109	80	243	180	157	105	146	181
	111	133	113	88	259	184	157	105	146	185
Han57	105	137	97	82	243	172	?	?	146	187
	107	137	109	90	263	182	?	?	158	193
Han58	107	133	103	80	243	182	155	99	146	185
	111	133	103	80	261	182	157	105	158	193
Han59	107	?	103	88	243	182	149	107	146	181
	111	?	117	88	259	182	159	111	158	185
Han60	107	133	103	80	243	184	?	?	150	?
	107	133	117	90	243	184	?	?	166	?
Han61	107	133	103	80	243	184	147	101	146	185
	113	133	103	90	243	184	153	105	158	185
Han62	107	133	103	80	243	182	?	?	146	?
	107	133	107	90	255	182	?	?	158	?
Han63	111	177	97	80	243	178	155	105	146	181
	111	177	115	90	259	182	157	109	146	187
Han64	107	177	97	80	?	180	155	103	146	185
	107	177	115	80	?	180	157	105	158	185
Han65	111	133	97	80	243	178	155	111	146	185
	111	133	105	90	259	186	157	111	148	185
Han66	107	133	97	80	249	?	159	111	146	185
	107	183	97	88	259	?	159	111	148	185

B.3.3 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Han67	107	133	97	80	259	178	?	105	146	185
	107	183	117	88	259	178	?	111	146	185
Han68	107	179	103	80	243	184	147	103	146	181
	107	183	117	82	253	186	157	105	146	185
Han69	107	133	107	88	243	182	157	111	146	185
	111	133	109	92	259	184	157	115	146	185
Han70	107	?	109	80	243	180	159	103	146	185
	111	?	117	80	249	180	159	105	148	185
Han71	107	133	103	82	255	178	155	107	146	?
	113	133	111	88	259	184	157	107	146	?
Han72	105	133	103	80	249	176	?	103	146	187
	107	133	111	88	259	178	?	109	148	201
Han73	111	133	105	90	243	180	147	105	?	185
	113	133	119	90	249	184	157	111	?	185
Han74	105	133	103	80	243	178	153	105	146	185
	107	133	103	80	249	180	153	105	158	191
Han75	111	129	103	80	243	176	?	101	146	181
	111	135	103	80	259	178	?	107	146	185
Han76	?	133	97	80	243	178	155	101	146	181
	?	179	115	80	265	184	157	107	158	185
Han77	109	133	103	80	?	178	153	107	136	181
	113	179	107	80	?	184	153	111	148	195
Han78	107	133	97	80	243	178	?	101	?	195
	111	133	103	80	249	184	?	105	?	195
Han79	?	133	?	80	243	176	147	?	146	?
	?	133	?	88	243	176	157	?	160	?
Han80	?	177	101	76	?	184	149	105	146	181
	?	177	103	90	?	184	153	107	146	185
Han81	105	133	99	80	243	?	149	105	146	181
	107	177	103	84	243	?	153	115	146	185
Han82	?	133	103	80	249	182	155	?	146	185
	?	135	107	80	249	182	157	?	148	189
Han83	103	135	99	80	243	178	153	105	146	185
	111	181	99	82	243	184	153	109	148	185
Han84	105	133	103	80	249	176	?	107	146	185
	107	179	111	90	249	180	?	113	146	185
Han86	?	127	97	90	243	?	153	?	146	185
	?	133	103	90	251	?	153	?	146	185
Han87	105	?	103	88	243	180	153	101	?	185
	107	?	103	90	249	180	153	109	?	187
Han89	?	133	?	80	243	178	147	105	136	185
	?	179	?	88	249	180	157	118	146	185
Han90	105	133	97	80	243	182	155	?	146	185
	111	133	109	88	259	182	157	?	146	185

B.3.3 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Han91	107	?	99	88	243	174	155	101	146	185
	111	?	107	90	259	180	157	105	146	193
Han92	107	?	97	80	243	182	147	?	146	185
	111	?	103	80	249	182	157	?	158	185
Han93	107	133	113	88	243	182	155	101	146	?
	111	181	115	88	243	182	157	105	160	?
Han94	?	133	103	80	243	178	147	103	146	185
	?	189	111	80	249	180	157	105	158	185
Han95	107	175	103	80	243	174	?	105	146	185
	109	183	109	88	249	178	?	107	146	197
Han96	107	129	115	80	243	176	155	?	142	195
	111	133	117	80	243	178	157	?	160	195
Han97	107	133	105	80	243	180	?	105	146	185
	107	133	107	80	249	182	?	111	146	187
Han98	107	133	?	80	243	172	?	101	146	185
	107	181	?	90	243	174	?	105	146	185

B.3.4 Hui autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Hui1	103	133	95	80	243	178	149	103	144	185
	103	133	95	88	243	186	157	103	144	195
Hui2	103	135	95	80	243	178	149	95	144	183
	107	179	95	80	249	186	157	103	158	195
Hui3	101	133	95	?	243	176	153	101	144	185
	107	175	115	?	243	184	153	105	144	185
Hui4	107	165	107	80	243	186	155	?	144	183
	107	179	107	80	243	186	157	?	144	183
Hui5	103	133	95	78	243	176	153	103	144	185
	107	167	107	78	249	186	157	109	158	195
Hui6	107	173	95	?	243	176	153	101	138	183
	107	133	105	?	243	176	157	101	158	183
Hui7	105	175	97	78	243	180	147	101	148	181
	109	175	97	78	259	180	155	101	156	181
Hui8	101	175	103	78	243	182	157	105	144	183
	107	175	115	84	259	182	173	109	158	181
Hui9	107	133	107	?	255	180	153	97	144	189
	107	133	107	?	257	180	157	105	144	201
Hui10	107	175	111	78	243	184	157	101	144	185
	107	183	115	84	243	184	175	109	144	187
Hui11	101	133	107	78	249	176	147	97	144	185
	101	133	107	86	249	180	147	97	144	185
Hui12	101	133	105	78	257	182	153	103	144	185
	101	133	109	86	259	182	157	103	158	185
Hui13	103	133	?	78	243	184	157	99	144	187
	103	133	?	78	243	184	157	105	144	187
Hui14	103	133	103	80	243	180	147	97	144	181
	107	133	103	90	243	186	157	97	158	181
Hui15	109	133	95	78	259	180	153	95	144	187
	109	133	105	86	259	180	153	101	156	195
Hui16	103	133	103	78	243	180	147	95	142	183
	109	133	103	78	243	180	153	95	156	183
Hui17	101	133	103	78	243	178	153	95	146	181
	101	133	103	86	243	182	157	103	158	187
Hui18	?	173	105	78	243	182	147	107	?	183
	?	173	105	78	243	182	159	107	?	183
Hui19	103	133	101	78	243	176	155	101	144	181
	107	177	101	84	243	184	155	115	158	187

B.3.4 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Hui20	107	133	101	90	243	176	155	103	146	183
	107	133	101	90	259	184	155	103	158	183
Hui21	103	133	101	78	243	184	149	103	158	185
	103	133	109	84	255	184	153	103	158	185
Hui22	101	133	105	78	243	182	153	95	144	185
	107	177	115	86	259	184	157	99	156	185
Hui23	103	127	101	80	243	172	153	105	144	183
	103	137	113	88	267	176	153	105	156	183
Hui24	101	135	103	78	249	184	155	107	144	183
	105	135	113	78	259	184	155	107	156	183
Hui25	103	135	95	80	243	176	153	115	144	187
	103	175	107	80	249	182	153	115	144	187
Hui26	101	135	103	80	243	180	147	103	144	185
	109	181	103	80	271	180	161	103	144	185
Hui27	103	?	103	78	243	180	155	115	144	185
	103	?	103	88	243	180	155	115	144	185
Hui28	103	175	95	78	239	180	147	105	144	185
	107	175	101	86	249	180	171	105	144	199
Hui29	107	?	97	80	243	182	147	97	158	183
	107	?	107	90	259	182	153	97	158	185
Hui30	103	133	93	78	249	182	153	103	144	183
	107	177	103	86	259	188	153	109	158	195
Hui31	103	133	105	86	243	182	145	105	144	183
	103	179	105	86	243	188	155	109	144	183
Hui32	101	133	107	78	249	186	157	101	142	185
	101	133	107	86	266	186	157	109	156	185
Hui33	103	133	95	84	243	180	147	95	142	183
	107	179	107	84	259	180	155	105	156	183
Hui34	111	133	97	90	243	178	147	103	156	181
	111	133	97	90	243	178	147	103	156	181
Hui35	103	133	105	?	243	178	147	105	156	183
	109	133	117	?	243	178	147	105	156	183
Hui36	103	133	99	78	243	184	147	103	144	183
	107	175	99	86	243	184	155	103	156	183
Hui37	103	133	95	78	243	180	147	105	146	181
	107	175	119	84	243	180	147	105	158	181
Hui38	101	133	105	78	243	176	155	95	144	181
	107	179	105	78	243	184	165	103	144	183
Hui39	101	135	97	80	243	178	155	103	158	181
	109	177	117	90	243	178	155	103	158	183
Hui40	111	133	101	80	243	178	155	95	144	183
	111	133	113	90	243	178	155	95	144	183
Hui41	103	133	101	78	243	178	147	103	144	183
	103	175	107	84	243	182	153	115	144	183

B.3.4 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Hui42	111	133	99	?	243	178	147	95	144	183
	111	133	99	?	243	178	147	109	144	183
Hui43	103	133	99	78	243	182	147	103	?	189
	103	177	107	84	243	186	153	103	?	189
Hui44	103	135	105	80	243	180	147	105	144	183
	109	135	111	80	263	180	147	105	156	183
Hui45	101	133	93	86	243	176	147	99	142	181
	101	133	107	86	266	182	147	99	142	193
Hui46	107	173	105	86	243	180	153	115	144	183
	107	173	111	86	243	180	153	115	150	189
Hui47	107	133	97	84	243	172	153	103	144	183
	107	133	103	88	243	172	153	103	144	189
Hui48	103	133	103	78	243	182	153	105	144	181
	107	173	103	86	243	182	167	105	144	181
Hui49	?	135	99	86	243	182	153	107	144	181
	?	175	115	86	243	182	159	107	144	183
Hui50	103	133	105	78	243	178	157	103	154	183
	103	179	105	86	243	178	157	109	154	183
Hui51	103	133	103	90	243	182	157	97	144	183
	103	133	103	90	243	182	157	101	158	183
Hui52	101	135	101	78	243	178	131	103	144	?
	101	179	107	86	243	184	131	103	158	?
Hui53	107	175	105	?	243	180	153	107	144	183
	107	175	105	?	249	184	159	107	160	183

B.3.5 Kucong autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Kucong1	105	133	109	82	243	178	157	101	144	181
	105	133	119	90	243	186	161	113	146	185
Kucong2	105	133	105	?	243	186	153	101	144	183
	105	133	111	?	243	188	173	107	144	185
Kucong4	107	133	107	?	243	178	157	107	146	181
	111	133	117	?	243	180	157	117	146	185
Kucong5	111	133	97	82	239	178	157	101	144	185
	111	133	111	90	243	182	159	113	146	187
Kucong6	107	133	107	82	243	178	157	99	144	181
	107	133	111	90	261	182	157	113	144	185
Kucong7	107	?	113	80	243	182	153	99	144	185
	111	?	115	90	249	184	173	109	160	185
Kucong8	105	?	111	90	243	184	147	95	146	185
	107	?	113	92	259	188	181	107	160	189
Kucong9	107	133	99	82	243	184	153	101	144	185
	107	137	107	90	243	186	157	115	160	185
Kucong10	107	131	99	90	239	184	157	101	144	187
	107	137	113	92	243	184	171	117	144	189
Kucong11	105	133	105	92	243	184	153	101	144	181
	107	137	119	92	257	186	157	115	144	185
Kucong12	103	133	97	82	241	178	157	101	144	187
	107	137	105	90	243	182	171	109	144	195
Kucong13	105	133	?	82	243	172	147	95	144	185
	107	133	?	90	243	184	153	107	144	185
Kucong14	105	133	?	82	243	184	157	95	146	181
	113	133	?	90	243	186	171	107	146	185
Kucong15	107	133	111	82	243	182	149	95	144	185
	111	135	113	92	249	186	157	107	144	189
Kucong16	105	133	97	82	243	184	157	101	144	185
	107	133	105	90	259	186	159	113	146	185
Kucong17	107	133	113	82	243	184	171	95	144	181
	107	137	115	82	243	186	171	107	146	185
Kucong18	107	133	105	92	243	178	147	107	144	185
	107	137	107	92	261	182	157	113	144	185
Kucong19	105	133	97	82	243	176	147	99	140	181
	111	135	105	90	243	184	157	109	144	185
Kucong20	107	133	103	82	243	178	157	99	140	181
	111	133	119	82	243	182	171	109	144	185
Kucong21	105	133	97	80	243	178	147	99	144	185
	107	135	103	92	243	186	171	107	144	185

B.3.5 cont

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Kucong22	105	133	107	82	243	182	147	107	140	185
	113	133	107	82	259	186	159	107	144	187
Kucong23	105	133	97	82	239	180	157	101	144	185
	107	177	113	90	243	184	171	123	164	189
Kucong24	103	133	103	80	243	178	153	97	144	183
	105	135	119	82	243	186	157	107	144	185
Kucong25	105	133	103	82	243	180	157	107	140	183
	113	175	111	82	243	186	157	109	144	185
Kucong26	105	133	97	82	243	182	157	?	144	181
	113	175	115	90	243	186	171	?	144	185
Kucong27	105	133	103	82	243	184	147	99	144	?
	113	175	117	82	243	186	157	107	144	?
Kucong28	105	133	99	?	243	178	153	99	144	183
	107	175	103	?	259	186	181	107	164	185
Kucong29	105	133	97	82	243	186	153	99	144	185
	107	175	103	86	259	186	159	117	158	189
Kucong30	105	133	103	90	243	180	153	97	140	181
	107	133	111	92	243	184	157	109	144	185
Kucong31	103	133	107	90	243	178	157	99	140	183
	107	133	113	92	243	180	171	109	144	185
Kucong32	107	137	99	90	243	182	153	97	144	181
	107	137	105	92	243	186	171	107	144	185
Kucong33	105	131	105	82	243	178	159	99	140	185
	107	135	113	92	243	180	171	113	144	185
Kucong34	105	133	109	82	243	178	147	99	140	183
	107	135	113	92	249	182	157	113	144	185
Kucong35	103	?	97	90	243	182	147	99	140	181
	111	?	119	92	243	186	171	113	144	185
Kucong36	111	133	103	82	243	180	153	101	144	185
	115	133	107	82	243	186	157	123	144	185
Kucong37	103	133	97	80	243	182	157	97	144	185
	107	133	113	82	259	188	157	107	160	187
Kucong38	105	175	107	82	239	184	157	101	144	185
	107	185	119	82	239	186	157	113	144	185
Kucong39	107	133	99	82	243	184	157	101	144	185
	109	137	103	92	243	186	157	113	160	185
Kucong40	107	135	105	82	243	182	147	101	144	185
	107	175	105	92	259	186	181	113	156	185

B.3.6 Miao autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Miao1	105	133	97	82	243	180	157	113	144	185
	107	177	99	90	243	186	171	113	146	189
Miao2	107	133	97	82	243	182	153	107	144	185
	109	179	99	90	265	186	157	107	156	193
Miao3	107	135	103	82	243	176	145	107	144	185
	109	137	113	90	259	184	153	113	146	185
Miao4	107	133	93	82	243	176	147	101	144	197
	109	181	97	90	259	178	157	107	160	199
Miao5	107	129	103	82	243	174	157	107	144	185
	109	133	111	88	259	182	157	109	144	197
Miao6	107	133	97	82	243	184	157	101	144	183
	113	133	103	82	243	186	157	107	158	185
Miao7	105	177	107	82	243	178	145	105	144	185
	111	185	115	90	245	182	153	107	164	187
Miao8	107	175	97	82	243	176	157	107	144	183
	109	179	107	82	245	184	157	123	160	185
Miao9	109	133	?	80	243	176	155	101	144	183
	109	133	?	82	259	186	157	105	144	185
Miao10	105	127	99	88	243	176	147	105	144	181
	109	133	105	90	243	182	157	107	160	185
Miao11	105	135	107	82	243	180	155	105	144	183
	109	177	109	90	259	186	157	107	144	185
Miao12	107	133	103	90	249	176	153	107	144	189
	111	133	109	92	259	186	157	109	144	197
Miao13	107	135	97	82	243	182	159	105	144	183
	111	135	111	82	259	186	159	107	146	185
Miao16	107	135	97	74	243	182	153	107	144	181
	107	179	107	82	259	182	157	107	144	185
Miao17	105	131	103	82	243	182	147	113	144	183
	107	135	117	92	245	186	157	123	160	187
Miao19	105	137	97	74	243	182	147	105	144	185
	111	185	111	82	243	182	159	107	144	195
Miao20	107	133	103	82	249	176	149	109	144	181
	111	177	103	92	249	184	157	111	160	189
Miao21	107	133	111	90	243	182	147	107	144	185
	107	179	113	92	245	184	157	107	156	197
Miao22	105	173	97	74	243	182	157	105	144	185
	105	183	107	82	261	186	161	107	144	187

B.3.6 cont

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Miao23	107	133	103	82	243	182	157	107	144	183
	109	135	107	90	261	184	161	107	144	185
Miao24	107	133	103	74	249	182	155	107	144	183
	107	135	107	82	259	184	157	107	160	185
Miao25	107	177	111	82	243	182	147	109	144	185
	109	179	113	90	245	186	157	109	144	185
Miao30	105	133	97	82	243	?	?	107	144	185
	107	183	115	90	263	?	?	107	146	185
Miao31	107	135	?	80	243	182	147	105	?	185
	111	135	?	88	263	186	157	107	?	185
Miao32	107	?	113	80	243	182	147	101	?	183
	109	?	115	88	243	186	157	109	?	185
Miao34	107	145	103	82	243	180	147	107	144	?
	107	145	107	90	249	186	157	107	156	?
Miao35	107	133	103	82	243	180	157	?	144	185
	107	145	107	92	263	186	157	?	160	185
Miao36	107	133	103	82	243	180	157	107	144	183
	107	137	109	92	243	186	157	113	144	185
Miao37	107	133	103	80	243	176	157	99	144	185
	109	175	115	88	243	178	157	103	158	195
Miao38	107	133	99	90	243	176	157	99	144	181
	113	133	109	92	259	178	157	103	160	185
Miao39	107	133	99	90	243	176	157	99	144	181
	113	133	109	92	259	178	157	103	160	185
Miao40	107	133	97	90	243	180	153	101	144	181
	109	133	111	92	259	182	159	101	144	193
Miao41	107	133	95	80	243	184	147	99	144	183
	107	133	97	84	243	184	161	115	144	185
Miao42	107	133	109	90	243	176	153	103	144	181
	109	133	119	92	243	178	157	107	158	185
Miao43	107	133	97	74	243	176	155	101	144	181
	109	187	103	82	243	182	157	109	158	183
Miao44	107	133	103	82	?	182	153	101	144	181
	111	135	113	92	?	182	157	105	160	197
Miao45	107	133	103	82	?	184	153	99	144	183
	109	175	103	90	?	184	171	107	160	187
Miao46	107	133	?	74	?	182	147	99	144	185
	107	183	?	82	?	182	157	107	160	185
Miao47	105	175	97	78	243	182	147	99	142	183
	105	185	107	82	243	182	157	107	146	187
Miao48	105	133	97	78	243	176	147	105	158	185
	109	177	105	82	249	180	171	107	160	199
Miao49	107	133	109	82	243	178	147	105	142	183
	109	185	111	92	243	182	155	105	146	187
Miao50	107	135	99	84	?	?	153	?	148	185
	109	179	109	92	?	?	153	?	159	197

B.3.7 Sala autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Sala1	103	129	105	106	245	174	157	109	144	185
	105	131	107	106	257	184	161	113	144	197
Sala2	103	129	103	78	245	180	147	105	144	183
	105	131	111	78	245	184	157	109	158	185
Sala3	103	125	103	80	243	180	147	107	144	181
	109	129	111	80	243	182	155	107	146	183
Sala4	103	129	97	80	243	180	147	109	144	185
	105	131	99	80	243	182	147	111	144	187
Sala5	105	129	103	80	241	176	153	101	144	183
	111	131	111	80	245	184	157	107	144	185
Sala6	109	129	109	82	243	174	145	101	144	183
	111	131	119	82	249	184	157	107	146	185
Sala7	103	129	75	82	241	180	147	107	144	183
	105	131	75	82	243	182	153	111	146	185
Sala8	109	171	75	82	241	174	147	109	144	183
	111	173	75	82	243	178	153	111	144	185
Sala9	103	155	75	82	241	174	147	107	144	183
	103	155	83	82	243	178	147	107	146	185
Sala10	103	129	101	80	243	176	147	103	144	183
	105	131	105	80	243	182	157	107	144	185
Sala11	103	127	85	78	243	176	153	103	144	181
	105	127	101	78	259	180	157	107	158	185
Sala12	103	129	103	78	243	180	157	107	144	183
	105	131	107	78	243	182	159	113	158	185
Sala13	101	129	107	82	243	176	153	105	138	183
	103	131	111	96	243	180	159	109	144	185
Sala14	107	129	107	84	243	184	157	105	144	181
	109	131	109	86	249	186	169	115	158	189
Sala15	105	133	97	84	241	182	149	107	144	185
	109	135	117	86	243	182	157	109	156	187
Sala16	111	129	105	80	245	178	147	107	144	183
	111	131	107	80	245	184	153	113	156	185
Sala17	107	175	85	104	243	178	145	107	138	183
	109	175	103	104	243	180	147	107	144	185
Sala18	105	129	97	78	243	178	149	107	138	185
	109	131	103	78	243	184	161	107	156	197
Sala19	105	129	97	78	243	178	153	107	144	185
	111	131	103	78	257	180	153	113	158	187
Sala20	111	131	103	86	243	172	153	105	144	185
	111	135	107	86	243	180	153	111	162	187

B.3.7 cont

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Sala21	105	129	111	78	243	176	149	103	144	185
	109	131	115	78	243	180	149	107	144	199
Sala22	103	173	103	82	243	182	147	?	144	185
	105	175	107	90	251	182	147	?	146	197
Sala23	105	129	107	78	243	180	161	107	144	185
	109	131	115	78	259	186	169	109	156	187
Sala24	103	129	107	78	243	182	147	?	144	183
	109	131	121	78	243	182	147	?	144	185
Sala25	103	173	103	78	243	178	157	99	144	181
	105	175	103	78	251	182	169	115	156	185
Sala26	103	129	87	82	243	170	147	101	144	193
	105	131	99	92	249	172	149	107	158	195
Sala27	103	125	87	78	243	182	153	109	144	185
	105	127	87	78	243	184	153	109	150	185
Sala28	103	129	87	78	243	174	153	105	144	181
	105	131	87	78	243	180	157	107	158	185
Sala30	105	171	109	78	243	182	?	107	144	189
	107	173	117	78	259	186	?	109	162	187
Sala31	103	?	101	78	243	182	155	105	144	181
	105	?	103	78	259	186	157	105	146	185
Sala32	103	129	99	84	249	182	157	105	146	181
	109	131	103	86	259	182	157	115	158	185
Sala33	109	175	111	78	243	176	155	109	144	187
	111	175	119	78	243	176	155	113	162	191
Sala34	105	129	103	78	243	182	147	99	144	181
	109	131	103	80	243	182	153	103	144	185
Sala35	103	131	97	78	241	180	157	97	144	183
	105	133	103	78	247	184	157	113	144	185
Sala36	105	129	103	86	247	182	147	97	158	183
	109	131	103	88	249	182	147	113	158	185
Sala37	103	167	93	78	243	176	157	107	144	183
	109	175	103	78	257	186	157	109	154	185
Sala38	105	129	113	78	243	176	155	107	144	187
	107	131	117	78	249	182	155	113	158	197
Sala39	105	129	99	84	243	182	157	107	144	185
	111	131	111	86	259	182	159	109	162	187
Sala40	109	175	107	78	243	172	147	107	144	183
	109	175	111	78	243	184	147	109	146	185
Sala41	105	129	109	76	241	178	149	113	?	185
	109	131	111	80	243	182	153	127	?	199
Sala42	105	133	97	78	243	172	155	113	144	183
	109	135	109	78	243	180	159	113	156	185
Sala43	105	129	105	78	243	182	147	107	158	183
	109	131	111	78	259	188	153	113	158	185

B.3.7 cont

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Sala44	105	129	107	78	245	184	147	107	144	185
	109	131	109	78	259	184	157	107	144	187
Sala45	103	129	97	86	243	182	153	107	142	185
	109	131	117	86	259	186	159	107	142	189
Sala46	103	129	103	78	243	176	153	107	142	193
	109	131	107	78	243	182	155	127	156	195
Sala47	101	125	103	78	241	180	157	103	144	181
	109	125	109	78	243	184	157	107	154	185
Sala48	103	131	99	84	243	178	147	107	144	185
	105	133	109	86	259	180	157	107	156	187
Sala49	105	129	97	78	241	180	147	103	142	185
	109	131	111	78	243	184	157	107	142	187
Sala50	105	131	?	92	243	?	155	105	?	185
	109	177	?	92	243	?	157	105	?	185
Sala51	103	131	97	92	255	176	147	101	144	185
	109	177	97	92	255	180	149	107	146	191
Sala52	103	131	103	82	243	180	147	95	158	185
	109	177	103	90	243	180	157	99	158	185
Sala53	103	131	105	90	243	184	153	105	142	185
	109	131	109	92	243	184	155	107	142	191
Sala54	103	133	107	82	243	182	147	109	142	185
	109	133	107	82	243	182	147	111	142	185
Sala55	103	131	103	82	243	?	145	99	142	185
	109	179	103	92	243	?	151	107	156	195
Sala58	103	129	95	90	243	184	151	107	142	181
	103	173	101	92	243	188	151	107	142	185
Sala59	103	179	?	78	243	182	151	105	?	185
	109	179	?	78	243	184	151	111	?	185
Sala60	103	131	?	90	243	180	145	105	?	?
	109	175	?	96	249	184	151	115	?	?
Sala61	109	131	109	82	243	180	145	115	?	?
	109	131	117	92	249	184	147	131	?	?
Sala62	105	131	107	82	243	186	155	101	142	185
	105	131	117	92	257	186	155	105	142	193
Sala63	107	131	95	82	243	182	?	107	142	185
	107	173	95	82	249	182	?	111	142	193
Sala64	99	131	97	82	243	182	?	105	?	181
	109	131	97	82	243	182	?	109	?	185
Sala65	109	131	95	82	243	180	155	105	156	181
	109	131	95	90	243	180	155	109	156	185
Sala66	103	127	103	82	243	176	?	113	142	181
	109	131	103	82	249	184	?	115	156	185
Sala67	105	177	?	82	243	170	147	107	142	185
	109	185	?	90	267	174	157	107	142	193

B.3.7 cont.

	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Sala68	109	131	91	82	243	180	?	107	142	185
	109	131	91	92	249	180	?	107	154	185
Sala69	103	131	79	82	243	168	147	107	140	185
	103	175	79	92	243	184	157	111	150	185
Sala70	101	131	?	82	243	?	?	107	?	185
	107	131	?	82	243	?	?	111	?	185
Sala71	103	131	105	84	249	168	147	107	144	185
	103	135	105	92	259	188	157	107	144	185
Sala72	105	131	99	88	243	180	157	107	144	181
	109	131	109	88	249	180	157	111	144	185
Sala73	103	131	103	82	243	180	155	113	154	181
	103	131	103	90	243	180	155	113	154	181
Sala74	103	131	103	82	251	176	155	105	156	183
	103	131	103	90	259	182	159	113	156	183
Sala75	105	175	97	90	243	180	155	105	142	185
	111	177	109	96	243	180	157	109	142	185
Sala76	103	175	111	82	255	182	153	101	142	181
	109	175	111	82	255	182	157	107	142	185
Sala77	109	131	91	82	243	174	147	107	144	185
	109	173	91	82	243	182	155	107	144	197
Sala78	103	131	109	82	243	178	?	101	140	187
	103	131	119	92	243	182	?	109	144	197

B.3.8 Tibetan autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Tibetan2	105	133	97	82	243	172	153	109	144	183
	109	133	101	82	243	182	169	113	158	185
Tibetan3	107	133	103	74	?	184	147	105	140	181
	113	133	103	82	?	186	157	107	144	185
Tibetan4	103	133	97	74	243	?	147	107	144	185
	111	133	103	82	261	?	155	113	144	185
Tibetan5	107	133	97	82	243	182	?	107	144	183
	111	133	111	82	249	186	?	107	146	185
Tibetan6	111	133	97	82	243	184	151	107	138	187
	111	133	97	90	261	188	157	111	144	187
Tibetan7	105	133	97	82	243	178	153	107	144	183
	107	175	103	90	243	184	155	113	158	185
Tibetan8	107	173	97	74	243	184	157	109	144	185
	107	175	107	82	243	188	157	109	144	187
Tibetan9	109	133	97	82	243	182	147	105	158	181
	111	133	103	90	243	188	147	107	164	185
Tibetan10	109	135	97	76	243	178	153	109	140	183
	111	171	97	86	243	178	153	109	162	185
Tibetan11	107	133	97	74	243	188	153	107	158	181
	109	173	111	82	243	188	157	115	160	185
Tibetan12	107	135	97	82	243	178	157	101	144	183
	107	177	97	82	243	186	171	109	156	185
Tibetan13	107	135	103	74	243	178	147	103	144	181
	109	177	107	82	243	184	147	109	148	185
Tibetan14	107	135	97	74	243	184	147	109	158	181
	111	177	105	82	243	188	157	115	158	185
Tibetan15	107	133	97	90	243	176	147	107	140	191
	111	135	97	92	243	186	157	107	144	193
Tibetan16	109	133	107	82	249	?	153	107	144	183
	109	133	109	82	249	?	157	113	152	185
Tibetan17	111	133	97	78	243	?	?	105	144	185
	111	133	103	90	243	?	?	113	144	191
Tibetan18	107	135	97	90	243	186	147	101	144	183
	107	173	103	90	249	186	157	107	144	185
Tibetan19	103	133	109	74	243	184	147	101	144	183
	105	175	109	82	243	186	157	107	144	185
Tibetan20	107	133	97	82	243	182	147	101	144	181
	111	173	97	90	243	184	147	107	144	185
Tibetan21	107	133	107	74	243	172	153	107	144	181
	111	133	109	82	243	188	153	109	158	185
Tibetan22	107	175	103	82	243	172	147	105	160	183
	111	177	107	82	243	184	167	111	160	185
Tibetan23	107	133	97	82	243	172	147	101	144	183
	111	175	99	92	243	188	157	109	158	185

B.3.9 Yao autosomal genotypes

Individuals	Autosomal Markers									
	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108	GABR3
Yao1	105	133	105	82	243	178	157	107	144	185
	111	133	109	90	261	182	157	117	144	185
Yao2	109	133	107	82	243	184	149	107	158	185
	109	133	113	90	249	186	159	109	158	185
Yao3	105	133	107	82	243	178	153	105	144	185
	105	163	115	90	245	184	157	113	144	185
Yao4	105	133	113	80	243	176	157	107	144	?
	111	133	115	90	249	184	159	107	156	?
Yao5	105	133	107	90	243	182	157	107	144	185
	107	163	117	92	243	184	157	113	144	199
Yao6	105	133	99	82	243	176	157	107	144	185
	109	133	111	82	261	182	157	113	144	185
Yao7	105	133	105	82	243	182	157	107	144	185
	105	135	115	92	261	182	157	107	144	197
Yao8	105	133	103	90	243	182	147	107	144	185
	105	133	103	90	243	184	157	117	144	185
Yao9	105	133	97	82	243	180	147	107	144	181
	107	133	111	90	249	182	159	107	144	197
Yao10	109	133	103	82	243	184	147	107	144	185
	109	133	105	90	245	186	153	107	156	189
Yao11	105	?	109	90	243	182	147	107	144	181
	109	?	115	92	243	182	147	109	146	185
Yao12	109	163	97	82	243	182	153	107	144	185
	111	163	97	82	243	184	159	109	156	185
Yao13	105	133	97	82	243	184	147	107	142	183
	115	133	97	90	243	186	153	113	144	185
Yao14	105	133	97	90	243	172	147	107	144	185
	107	133	107	90	261	184	153	113	144	185