2001

Genetic Diversity in the Bo'an, Salar, and Dongxiang : Co-Resident Muslim Populations in Gansu Province, P.R. China

Thomas Baric

USE OF THESIS

Genetic diversity in the Bo'an, Salar, and Dongxiang: co-resident Muslim populations in Gansu Province, P.R. China.

by

Thomas Baric

Student number: 0970615

A thesis submitted in partial fulfilment of the requirements for the award of

Bachelor of Science (Human Biology) with Honours

Faculty of Communications, Health and Science, Edith Cowan University

Abstract

Patterns of genetic diversity within and between three co-resident Muslim populations from Gansu Province in the Peoples Republic of China were examined and the results contrasted with historical information. This study of members of the Bo'an, Salar and Dongxiang communities will contribute to a clearer understanding of the origins and migratory patterns of Muslims in PR China, and more generally the effect of population subdivision on gene pool structure and composition. Ten autosomal and five Y-chromosome microsatellite loci were genotyped to determine allele distribution patterns. Subsequently, the D-loop region of mitochondrial DNA was sequenced to complement the autosomal and Y-chromosome data. To infer between- and within-population relationships, data were analysed by several alternative statistical techniques based on either the Infinite Allele Mutation model or the Stepwise Mutation model. Due to the endogamous nature of the three populations, increased levels of homozygosity at autosomal loci were observed. Y-chromosome data exhibited major differences between the study populations, whilst mitochondrial DNA suggested more consistent inter-community relationships. Demographic information was also assessed to provide a more detailed account of population structure and phylogeny.

Declaration

I certify that this thesis does not, to the best of my knowledge and belief:

i) incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education;

ii) contain any material previously published or written by another person except were due reference is made in the text; or

iii) contain any defamatory material.

Signature...... [redacted]

Date.........25/01/01.........

Acknowledgements

First and foremost, I would sincerely like to thank my supervisor Professor Alan Bittles for his guidance, encouragement and support, never have I learnt so much about the world in such a short period of time. My many thanks also go to Dr. Wei Wang without whom this project would not have been possible, and Salvatore Di Grande who I cannot thank enough for his friendship, help and expertise throughout this study.

Thanks also to Dr. Peter Roberts and Dr. Angus Stewart at the Department of Human Biology at ECU for your support and advice over the past few years. Karen Downes for her cheerful assistance and wonderful organisational skills. A special thankyou to everyone at the Centre for Human Genetics for his or her help during the year. Thanks also go to anyone whom I haven't mentioned personally but who helped along the way.

Finally, thankyou to my ever supportive Mother and a wonderfully patient Fiancée, Lissy.

Table of Contents

Chapter 1

Introduction

Chapter 2

Historical Background

Chapter 3

Subjects and Methods

Chapter 4

Results

Chapter 5

Discussion

Appendices

Appendix A

Appendix B

Appendix C

Appendix D

Appendix E

Appendix F

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Watson and Crick published their analysis of the DNA double helix in 1953 and thus set in motion a revolution in biology that has culminated in the Human Genome Project (HGP). Officially launched in the United States of America on the 1st October 1990, the HGP was established as a 15-year scientific venture coordinated by the Department of Energy and the National Institutes of Health (NIH) (Venter *et al.*, 1998). The HGP is an international collaborative effort administered by the Human Genome Organisation (HUGO), with 18 member countries including Australia. The focus of the HGP project was to identify the estimated 50,000 to 100,000 human genes by sequencing over 3 billion base pairs of the entire genome. Besides the furtherance of scientific knowledge, the fundamental rationale of the HGP is to acquire information about our genetic constitution and the role of different genes in health and disease.

An additional consequence of the HGP was the establishment of the Human Genome Diversity Project (HGDP) in September 1993. Affiliated with HUGO, the HGDP planned to build on the many technical advances made in the HGP. The primary goal of the HGD Project is to explore the full range of genetic variation of the humans worldwide. This scientific venture is designed to collect information on human genome variation in order to understand genetic variation within and between human populations. The information obtained by the HGDP will be used to improve our knowledge of human biological history and the biological relationships among different human groups. It may also contribute to our understanding of the causes of particular human diseases (Resnik, 1999).

Nucleotide sequence variations identified in the HGDP from autosomal and Y-chromosome markers and mitochondrial DNA (mtDNA) data can be used as genetic indicators and may possibly contribute to variations in disease susceptibility. For example, common forms of variation or polymorphism include those that are responsible for the different blood groups. The polymorphic distributions of various alleles can play an important role in quantitative traits such as drug sensitivity and behavioural phenotypes. These polymorphisms would be evident on an individual basis, within populations and between populations. From a population genetic perspective the data would also be used to contribute to our understanding of human migratory history.

1.2 Purpose of the study

The present investigation aims to compare and contrast allele polymorphisms in three Muslim communities in the Peoples Republic of China (PR China), the Bo'an, Salar, and Dongxiang. In conjunction with the available historical and anthropological information, genetic data will be used to examine the effects of population sub-division on genetic diversity in each of the populations. The study will attempt to compare genetic diversity at autosomal and Y-chromosome loci, and in the D-loop region of mitochondrial DNA.

1.3 Significance of the study

The three Chinese Muslim communities chosen for the study are excellent models for the study of human genetic diversity and evolution, since they share common characteristics. These include:

1. Common residence in Jishisan County, Gansu Province;
2. The Muslim religion;
3. Community endogamy.

In contrast, the community have distinctive languages, two of which are Mongolian-based and one Turkic in origin. None of the languages has a written script.

The investigation also contributes to a larger body of work being conducted by Prof. A.H. Bittles and Dr. W. Wang on the comparative analysis of genetic diversity between Indo-Pakistani and Chinese minorities.

1.4 Research questions

1. Can distinct patterns of genetic diversity be defined for the three study populations using a panel of established autosomal and Y-chromosome microsatellites, and can the investigation of an established mtDNA sequence complement these results?
2. Are comparable patterns of diversity observed with autosomal and Y-chromosome microsatellite markers and mtDNA?
3. Can patterned variations at Y-chromosome loci be associated with male-directed gene flow and founder effect?
4. Can patterned variations in the mtDNA sequence be associated with female-directed admixture?
5. Are the results of the study compatible with the historical and ethnographic evidence from the three populations?

Chapter 2

Historical Background

2.1 Introduction

The proposed study aims to compare molecular genetic data with historical information on the origins of the Bo'an, Salar, and Dongxiang populations of Jishisan County, Gansu Province in the PR China. As a consequence, a basic understanding of the demographic and social aspects of these populations is required to complement the theoretical and empirical information derived from genome-based studies.

2.2 The Peoples Republic of China (PR China)

The population of PR China is the largest in the world and currently exceeds 1,200 million people. Continental China has a surface area of some 9,597 million square kilometres and is divided into 22 provinces, three municipalities (Beijing, Shanghai and Tianjin), and five autonomous regions (Family Planning Commission, 1998; Roberts, 1999). The coastline, bordered by the Yellow, East China, and South China seas, is about 12,000 km long. China also shares land borders of 21,260 km with 14 other countries: North Korea, Russia, Mongolia, Kazakhstan, Kyrgyzstan, Tajikistan, Afghanistan, Pakistan, India, Nepal, Bhutan, Burma, Laos, and Vietnam.

On a global scale the world population is estimated to be 6067 million people, which suggests that approximately one in every five people on the planet are Chinese (PRB, 2000). Within PR China 92% of the population are Han. The term Han suggests a monoethnic people, however this is by no means an accurate description for China's majority population (Gladney, 1998). In addition to the majority Han there are 55 ethnic minorities, comprising approximately 91 million people (Family Planning

Commission, 1998). Thus modern China is an ethnically diverse nation with many cultural, geographic and linguistic differences.

The government of PR China officially recognises ethnicity or *minzu* (nationality) within the country on the basis of specific population characteristics, such as a common territory, religion, language and economy (Gladney, 1998). The Salar, Bo'an and Dongxiang form part of the ten Muslim *minzu*, the others being the Uygur, Kazak, Kirghiz, Tadjik, Uzbek, Tatar and Hui. Collectively, the ten Muslim *minzu* form 1.3% of the total Muslim minority population in China (Family Planning Commission, 1998).

For centuries, virtually all foreigners entering China originated from less developed societies along their land borders, a circumstance which conditioned the Chinese view of the outside world. The Chinese saw their domain as the self-sufficient centre of the universe and derived from this image the traditional (and still used) Chinese name for their country *Zhongguo*, literally translated as 'Middle Kingdom' or 'Central Nation'. China perceived itself to be surrounded on all sides by *wayji,* or 'outside barbarians', whose cultures were demonstrably inferior by Chinese standards. Minorities who lived within the country were referred to as *neiyi,* or 'inside barbarians' (Gladney, 1998).

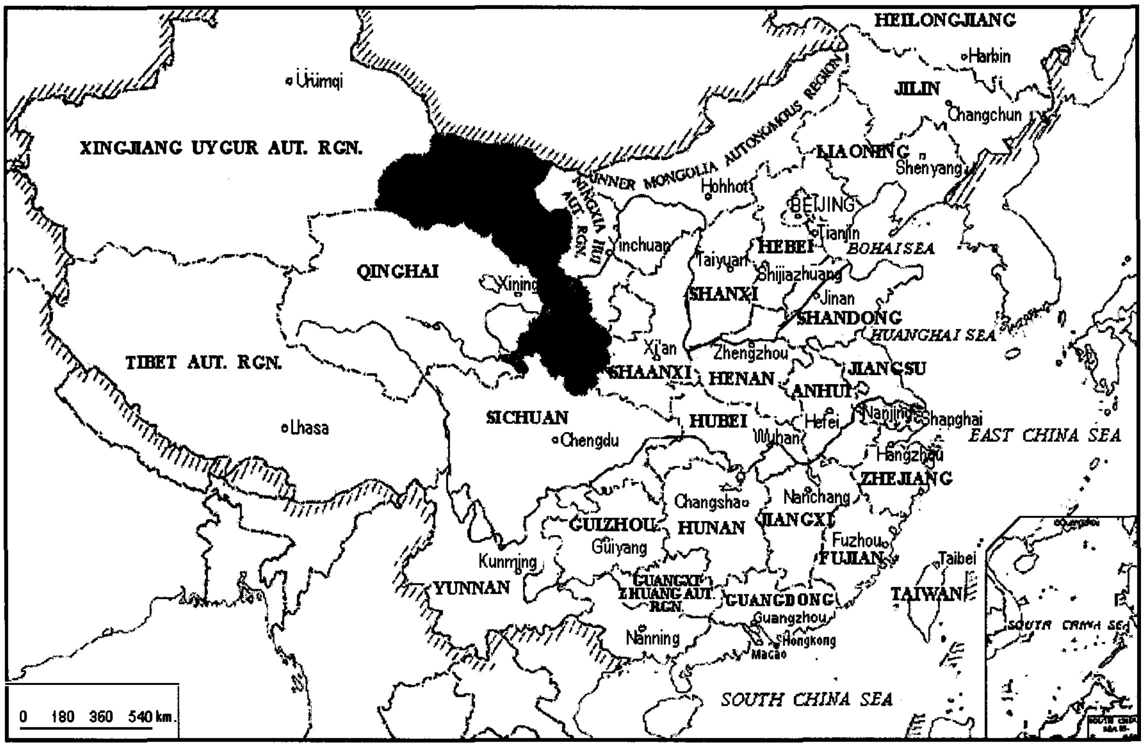
The threat of barbarian invasion of the 'Middle Kingdom' was a recurrent theme throughout Chinese history, as was the assimilation of the Chinese with non-Chinese. In the thirteenth century AD, Mongols from the northern steppes became the first foreigners to conquer China. Although not as culturally developed as the Chinese, the Mongols left some imprint on Chinese civilisation, for example the Bo'an and Dongxiang in this study still use Mongolian as their language, at the same time heightening Chinese perceptions of a threat from the north. In the mid-seventeenth century China came under foreign rule for the second time, and again the Manchu invaders originated from the north and northeast of the country. The process of assimilation continued over the centuries through conquest and colonisation until what is now known as 'China Proper' was brought under unified rule by the communists during the Peoples Republic Period (1949 AD- present) (Roberts, 1999).

2.3 Gansu Province

The Province of Gansu is located in the mid-northwest of the PR China. Positioned on the upper reaches of the Yellow River (*Huang He*) and with a predominant landscape of mountains and plateaus, Gansu has a corridor that extends north and west between the Quinghai-Tibet plateau in the south and the Gobi desert in the north. During the Ch'in dynasty (221-206 BC), Gansu was first officially recognised as under Chinese administration. Since it served as a passage for trade along the Silk Road between Eastern China and Europe, Gansu became an important strategic outpost and communication link for the area (Du and Yip, 1993).

Following the establishment of PR China in 1949, autonomous regions were set up for Muslim communities within Gansu. A significant factor for official recognition as an ethnic minority was a culturally specific language (Gladney, 1998). As the Bo'an, the Salar and the Dongxiang each possess their own distinct language, they were accorded official recognition as ethnic minorities and were granted autonomous status in Jishisan County, Gansu (Du and Yip, 1993) (fig. 2.1). Other numerically smaller Muslim communities who did not possess a culturally identifiable language were grouped within the larger Chinese Hui Muslim minority, rather than being recognised as minorities in their own right (Gladney, 1998).

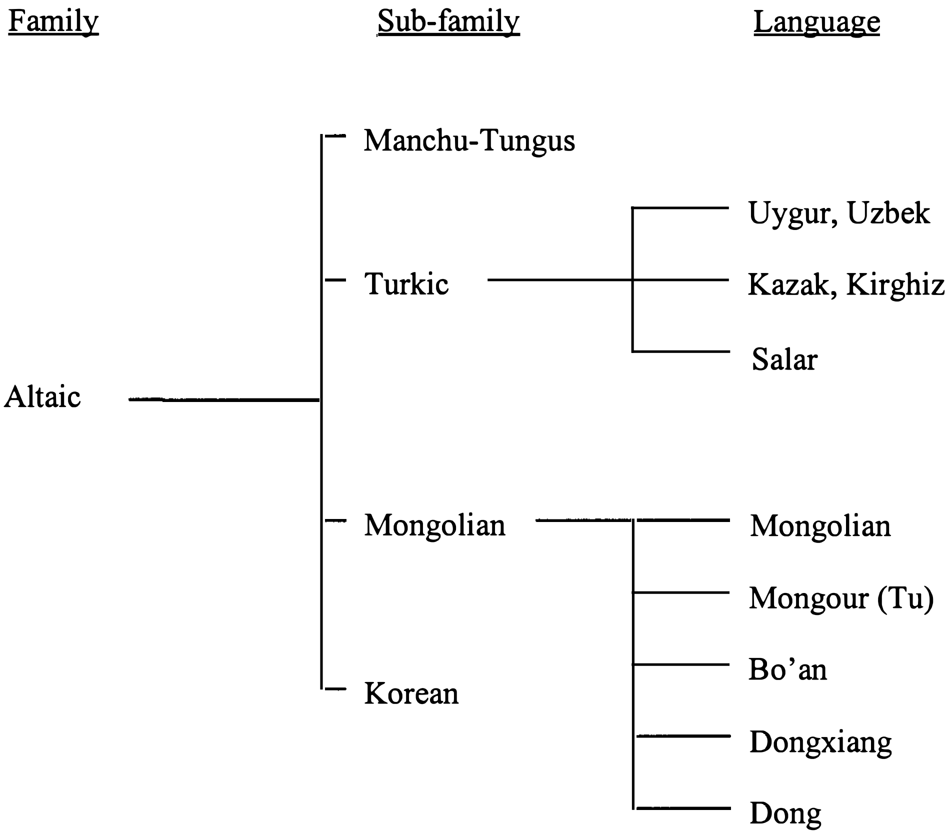
Figure 2.1 Map of the Peoples Republic of China indicating Gansu Province



The four main language families in China are Altaic, Sino-Tibetan, Mon-Khmer and Indo-European (Gladney, 1998). As illustrated in Figure 2.2, the Bo'an, Dongxiang and Salar languages all belong to the Altaic group of languages. Although the spoken language of the Bo'an derives from the Mongolian group, there also has been considerable Han influence on both the spoken and written forms of the language, and it has been estimated that over 40% of words may be of Han origin. As a consequence of this strong Han influence, the Bo'an use Han characters since they do not have their own script (Du and Yip, 1993).

The Dongxiang language also is derived from the Mongolian group of the Altaic family. The majority of Dongxiang can speak the Han language and write using the Han script. In addition, many Dongxiang can write and spell in Arabic, and the Muslim religious faith is strong within the community. The Dongxiang refer to themselves as Sarta, literally translated as "the Islam Se Mu" (coloured eyes) from Central Asia (Du and Yip, 1993).

Figure 2.2 Altaic language family of China



(Adapted from Gladney, 1998)

In common with the Uygur, Uzbek, Kazak and Kirghiz, the Salar language stems from the Turkic sub-family of the Altaic language family. The predominant influences on Salar in recent times have been the Han and Tibetan languages, and because of a lack of their own written language the Salar have adopted the Han script. Prior to 1949 over 97% of the Salar people were estimated to be illiterate (Du and Yip, 1993).

2.4 Muslim Origins in China

The Chinese political pattern has been described in historical terms as a series of dynasties, one following another in a cycle of ascent, achievement, decay, and rebirth under a new family. Beginning with the Western Han Dynasties, Table 2.1 shows the major Chinese dynasties/periods of the modern era in chronological order.

Table 2.1 Chinese dynasties

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
Sung 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	1908 AD – 1949 AD
Peoples Republic Period	1949 AD – present

Recorded Chinese Muslim history dates back some 1400 years to the Tang Dynasty (618-907 AD), and suggests that followers of the Islamic faith variously entered China as troops, merchants and political emissaries from Arabia and Persia. Groups and individuals subsequently settled among the various townships throughout China, especially along the Silk Road, and contributed appreciably towards the local and national economies (Wong and Dajani, 1988). According to Chinese-Islamic folklore, the first Muslim emissary from Arabia arrived in the 7th century AD. Saad ibn Abi Waqqas, a devout follower, uncle, and companion to the Prophet Muhammad, led a delegation to China to meet the Emperor Yung-Wei (Rahman, 1997). Saad brought gifts for the Chinese Emperor and attempted to convert the Emperor to Islam. Although the attempt was unsuccessful, Yung-Wei respected the teachings of Islam and judged that they were compatible with Confucianism. As a consequence, the Emperor granted Saad freedom to spread the Islamic faith among his people, resulting in the establishment of the religion in China.

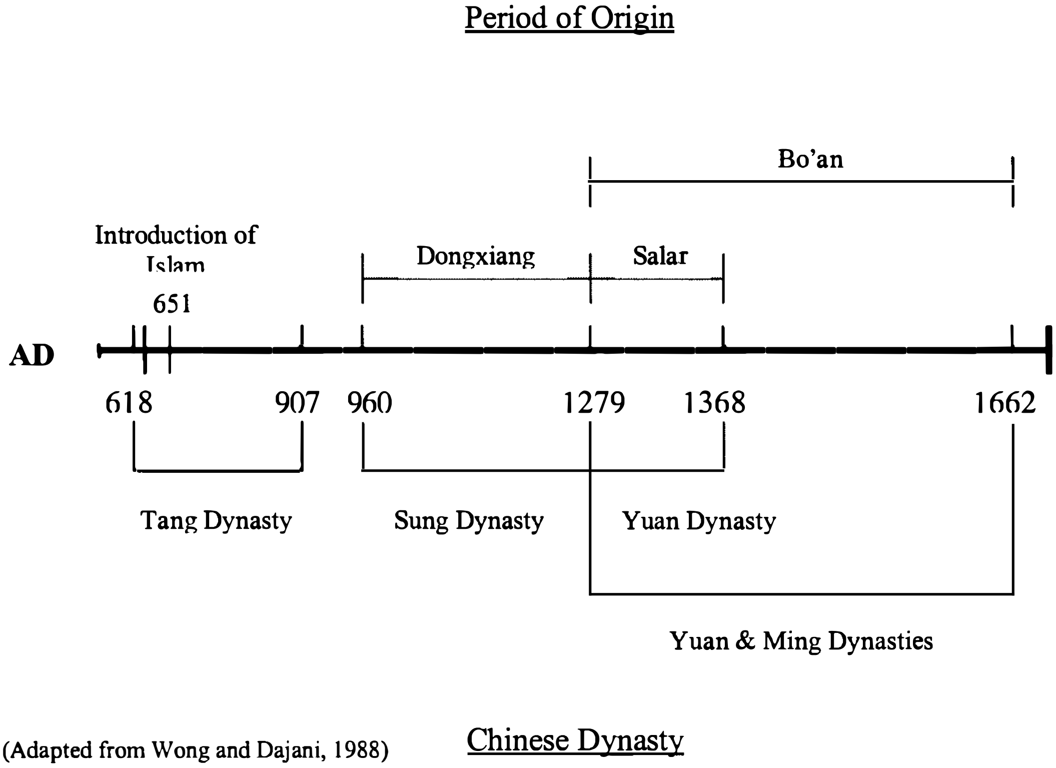
For a more historical perspective, the Tang Annals (618-907 AD) recorded the introduction of Islam into China in 651 AD, following the arrival of an emissary sent by the King of Arabia (the third caliph, Uthman) to gain favour with the Chinese and allow the free movement of trade (Wong and Dajani, 1988). Over the next hundred years, dealings between China and Arabia resulted in cultural exchanges. The Chinese were however threatened by the contrast in cultures, resulting in increased tension between the two nations.

The Tang Annals in 751 AD describe the Battle of Talas in terms of this cultural disparity (Wong and Dajani, 1988). During the battle Chinese and Arab armies met for the first time, and the defeat of the Chinese paved the way for Islamic control of central

Asia. The end result was the creation of a link between the two empires, China and Arabia, with two of the most powerful and heavily populated cities in the world at its western and eastern ends, Baghdad (City of Peace), and Ch'ang-an (City of Eternal Peace), the capital of the Tang Dynasty.

Only four years after the Arabs had defeated the Chinese at Talas, a Chinese military commander, Lu-Shan, rebelled against the Tang emperor and seized control of Ch'ang-an. The Chinese Emperor Su Tsung appealed to the second Abbasid Caliph, Abu Ja'far al-Mansur to help in the recapture of Ch'ang-an. The Caliph agreed and sent an army of 4000 troops, which successfully recaptured the city in 755 AD. As illustrated in figure 2.3, many of these troops remained in China and married local women, thus contributing to the formation of Muslim minorities (Wong and Dajani, 1988).

Figure 2.3 Timeline summarising the formation of the Bo'an, Dongxiang and Salar communities.



(Adapted from Wong and Dajani, 1988)

The establishment of the Yuan Dynasty (1279-1368 AD) followed the destruction of the preceding Sung Dynasty (906-1279 AD) by Genghis Khan and the consequent establishment of Mongol rule over much of China. The Mongols resettled some Islamic groups to important administrative posts in China, thus facilitating the spread of Islam inwards toward central China and to Gansu Province. During the period of the Mongol reign the Muslims were second in influence only to the Mongols, and they became an important part of the ruling bureaucracy (Wong and Dajani, 1988).

It was during the Ming Dynasty (1368–1662 AD) that the separate alien status of the Muslims changed. Assimilation was favoured and individual customs, languages, and traditions became obsolete under the reign of Emperor Hongwu (1368-1398). During the fifth year of the Emperor's reign an imperial edict was issued that stated:

> *"Mongolians and Se Mu (Hui) people are allowed to marry Chinese but not their own kind" (Du and Yip, 1993).*

Under the Ming Dynasty, Muslims were to become fully integrated into Han society (Wong and Dajani, 1988). This resulted in many Muslim men taking the name of their Han wife or using Chinese characters that closely resembled their own name, such as Hu for Hussein and Sai for Said (Rahman, 1997).

With the establishment of the Qing Dynasty (1662–1908 AD) the integration policy for Muslims again changed. Various Muslim uprisings against the Manchu rulers, including those in Kashgar (1758-1759 AD) and Quinghai (1781-1784 AD) prompted the establishment of imperial policy relating only to Muslims. Special laws assembled Muslims into two distinct groups, the loyalists who supported the Qing Dynasty and those who did not. Local populations who sided with the Manchu rulers were allowed to retain their own religious leaders, follow their own dietary rules, and not wear the queue (a distinctive hair cut symbolising Manchu acquiescence). In effect, these laws were an attempt to allow the Qing Dynasty to expand its territorial reach and at the same time suppress uprisings (Ebrey, 1999).

Muslim life during the Qing period was mostly focused around communal life at the local mosque and the *Ahong,* or teacher, who was appointed by the community elders (Ebrey, 1999). This style of Islam centred amongst the local village people is known as *Gedimu*, and is still in practice (Gladney, 1996). The arrival of Sufism (Islamic mysticism) in China caused many battles between different Muslim orders, and clashes between Muslims and the Imperial army. Centred on a Sufi or teacher, there were four main *menhuan* (or saintly lineages), the Jahriyya, Khufiyya, Qadiriyya and Kubrawiyya. Each *menhuan* was based on their own interpretation of Islamic texts and ideological beliefs (Gladney, 1996).

The impact of Sufism or Muslim Chinese was highlighted by battles that took place between the Qing imperial forces and the armies of the Sultanate of Dali, in which an estimated 10 million persons died between 1855 and 1873 (Ebrey, 1999). Since the present total population of Muslims in China is approximately 16 million (Family Planning Commission, 1998), and assuming that around half of the casualties were supporters of the Sultanate, i.e., Chinese Muslims, loss of life on this scale may have had a significant bottleneck effect on the present day population structure of the Salar, Bo'an, and Dongxiang.

Communist rule was established in 1949 with the creation of the Peoples Republic. To strengthen support for the government in distant regions of China a decision was made to afford autonomy to ethnic minorities. This step had been presaged in the Communist Party Constitution of 1932, which allowed for complete autonomy of various minority regions, prefectures and counties (Gladney, 1996). The PR China government recruited anthropologists and demographers to survey the population of China. As a result communities who previously were collectively referred to as Hui were divided into 10 separate Muslim *minzu*, the Uygur, Kazak, Dongxiang, Kirghiz, Salar, Tadjik, Uzbek, Bo'an, Tatar, and Hui. The first nine were given separate *minzu* status as they each had a unique language of their own. The reformed Hui *minzu* comprised those Muslims who did not have a language of their own but who spoke the dialects of the surrounding populations (Gladney, 1996).

2.5 The Salar

In 1997 the Salar population numbered 87,546 (Family Planning Commission, 1998). Approximately 80% of the Salar live within the Salar autonomous region, situated in Xunhua County in eastern Quinghai Province. The remainder are located throughout the neighbouring Hualong County, and in the Bo'an-Dongxiang-Salar Autonomous County of Jishishan, Gansu Province.

The origin of the Salar people can be traced back to the Yuan Dynasty (1279-1368 AD). With Mongol rule established over many parts of China, Muslims, including the founders of the Salar community, were relocated to important administrative posts (Wong and Dajani, 1988). It is suggested that 170 Salar families embarked on a journey along a trade route beginning in the city of Samarkand in Uzbekistan, Central Asia. Through time this journey took them via Turpan and Suzhou, and eventually led them to settle in Xunhua County (Du and Yip, 1993).

By tradition the Salar are farmers, and they are devout Muslims. From the time of the Yuan Dynasty (1279-1368) through to the beginning of PR China (1949), the

Imams, acting in concert with community chiefs and village headmen, enforced strict adherence to religious scriptures via a near-feudal system. In 1949 this system was abolished and the Salar community was granted autonomy in Jishisan County, Gansu Province (Du and Yip, 1993).

The Salar people are distinguishable from the Han by their tall stature, high nose bridge, deep-set eyes and facial hair. Another physical characteristic is congenital colour blindness, said to effect approximately 6% of Salar males (Du and Yip, 1993).

2.6 The Bo'an

The Bo'an or Bao'an are one of the smallest of the officially recognised ethnic minorities in China, with a total recorded population in 1996 of 11,683 persons (Family Planning Commission, 1998). The term Bo'an actually refers to a specific location, a fortress in Tongren County, Quinghai Province, which was built to protect its inhabitants toward the end of the 16th century. Translated, Bo'an literally means "security" (Du and Yip, 1993).

The first recorded evidence of the Bo'an arose during the Yuan and Ming Dynasties (1279-1662 AD), although it has been suggested that the Bo'an ethnic group had developed gradually over many years. Mongolian forces occupied and settled in Tongren County, and along with neighbouring communities including the Hui, Tibetans and the Tus, they intermarried and gradually became assimilated. During this period Bo'an customs and cultural identities were shaped, including acceptance of the Islamic religion (Du and Yip, 1993).

During the Yuan and Ming centuries (1279-1662 AD) the Islamic beliefs of the Bo'an led to their oppression and exploitation. With the establishment of PR China in

1949, the Bo'an and other ethnic minorities (including the Dongxiang and Salar) were granted autonomy. In present-day PR China the Bo'an primarily are employed in agriculture (Du and Yip, 1993).

Due to the small numbers of Bo'an, more than 20% of marital unions are with surrounding ethnic groups. As a result, their physical features are scarcely distinguishable from the local Hui, Salar and Dongxiang communities (Du and Yip, 1993).

2.7 The Dongxiang

The Dongxiang ethnic minority comprises over 373,600 individuals, of whom more than 50% are resident in Dongxiang Autonomous County in northwestern Gansu Province, with the remainder mostly living in Linxia and Hezheng Counties, Gansu Province (Family Planning Commission, 1998).

Several theories have been put forward as to the origin of the Dongxiang people and their culture. One theory suggests that at some point in his westward advance, Genghis Khan (1162-1227 AD) stationed troops in the Dongxiang area. These Muslim troops then married the local women, worked the land, and the Dongxiang are the descendants of the resultant progeny. Another hypothesis suggests that the Dongxiang are a mixture of various ethnic groups, including the Han, the Hui, and the Mongolian and the Tibetan peoples (Du and Yip, 1993).

The strength of their Islamic faith is evidenced by the fact that at one time there were more than 595 mosques in the Dongxiang area, which meant that a place of worship was available for every 30 households. Muslims in the area were divided into three sects, the old, new and emerging sects. Constant feuding among these sects led to a policy of divide and rule, causing a great deal of disunity in the Dongxiang region. This disharmony among the Muslim Dongxiang was alleviated following the establishment of the PR China (Du and Yip, 1993).

The general physical characteristics of the Dongxiang include a fair complexion and light-coloured eyes, with some individuals possessing light blue eyes. The men have elliptical faces with high noses, deep eye sockets and thick beards (Du and Yip, 1993).

Chapter 3

Subjects and Methods

3.1 Subjects

The blood samples used in the study were collected in 1999 by Dr. W. Wang and colleagues from the Institute of Genetics of the Chinese Academy of Science. Informed consent was initially obtained from community elders, with individual subject compliance given on a voluntary basis. Finger-prick blood samples were obtained from the subjects using sterile lancets. The samples were taken on to 3MM™ Whatman filter paper on location in Gansu Province, air dried at room temperature, and sent by air mail to The Centre for Human Genetics at Edith Cowan University where they were stored at -80°C. The study was based on randomly selected and unrelated samples from 67 Bo'an (male 47/female 20), 64 Dongxiang (male 49/female 15) and 81 Salar (male 52/female 29), subjects.

3.2 Microsatellite markers and genetic diversity

3.2.1 Microsatellites

Microsatellites are defined as a sequence of short tandem DNA repeats, often between 1 and 6 base pairs in length, which are frequently characterised by the existence of numerous alleles. Also known as short tandem repeat polymorphisms (STRs), these sequences are highly informative, abundant and randomly distributed throughout the human genome. Microsatellites are inherited in a Mendelian fashion, and they are members of a larger group of repetitive DNA sequences that include satellite DNA, minisatellite DNA and transposable elements (Charlesworth *et al.*, 1994; Strachan and Read, 1999).

As indicated in table 3.1, microsatellites have three basic forms - pure, compound and interrupted. These three forms of microsatellite can be expressed interchangeably, in either a single form or a combination of forms (Jarne and Lagoda, 1996).

Table 3.1 Microsatellite classification

Pure	CACACACACACACACACACACACACACA
Compound	CACACACACACA**GAGAGAGAGAGAGA**
Interrupted	CACA**TT**CACACA**TT**CACACA**TT**CA**TT**CA

The characteristic high mutation levels and locus hypervariability exhibited by microsatellites make them useful tools in the analysis of closely related populations. Each allele of a particular microsatellite locus in the human genome is theoretically neutral in effect. This would suggest that microsatellites have a constant rate of evolution, which is independent of population size (Esteban *et al.*, 1999; Kimura, 1968).

3.2.2 Microsatellite mutation

Single microsatellite mutation rates have been estimated as between 1000 and 100,000 per gamete, thus indicating the probability of high levels of polymorphism (Murray, 1996). There is evidence that both microsatellite repeat length and base composition contribute toward the mutation process. For instance, dinucleotide repeats are believed to mutate faster than tri- and tetranucleotide repeats, and sequences with a high AT content mutate faster than sequences with a high GC content (Murray, 1996). Two mechanisms have been proposed for the mutation process, i.e. polymerase strand slippage at replication, and unequal crossing-over at meiosis. Although these mechanisms are not yet fully understood, it has been suggested that they act in concert. Polymerase strand slippage results when a newly synthesised strand of DNA fails to follow normal base pairing while disassociating itself from the polymerase complex. This results in a transient bulge in the newly synthesised strands of DNA that are either repaired by enzyme-mediated systems of repair, or ignored. The outcomes of both these mechanisms result in STR polymorphisms (Jarne and Lagoda, 1996).

3.2.3 Models of microsatellite evolution

The mechanisms of microsatellite mutation have led to the development of two major models that seek to explain their consequences, the infinite allele model (IAM) and the stepwise mutation model (SMM). The models relate to parameters used to estimate population differentiation and distance, and thus are of potential importance in population genetics. The IAM assumes that, within a given population, each mutation gives rise to a unique allele and is therefore equivalent and identical by descent (Kimura and Weiss, 1964). The SMM assumes that a mutation changes the allele by one unit, and so infers that alleles of the same size are more closely related but not necessarily identical by descent (Shriver *et al.*, 1993; Valdes *et al.*, 1993).

However, to clearly appreciate the divergence between populations and place the IAM and SMM models into perspective, the phenomenon of homoplasy needs to be addressed. The term homoplasy refers to the occurrence of a mutational event that changes the allele size and sequence only to revert the allele back to the original state. Microsatellite mutation does not rely solely on the gain or loss of a single repeat unit; rather the mutation process can involve the repeat of several units at a time (Di Rienzo *et al.*, 1994; Weber and Wong, 1993). If two alleles are identical by descent, they have been inherited without mutation from the same ancestral allele. Alternatively, if an allele is inherited with identical size and sequence and has undergone homoplasy then it is said to be identical in state. Homoplasy also refers to an allele that may be different in size and sequence at protein level, but remains functionally equivalent to the ancestral allele.

3.2.4 Genetic distance

The measure of genetic distance between populations is a quantitative value that seeks to assess, in relative terms, the length of time passed since the populations existed as a single cohesive unit (Nei, 1987). Polymorphisms derived from microsatellite alleles are commonly utilised to evaluate genetic distance. Microsatellite polymorphisms are usually correlated in terms of distance to time (typically in generations). In population genetics this technique has led to a clearer understanding of human evolution, and as such it forms a major focus of the present study.

Essentially, distance measures can be sub-divided into two groups, used either in population studies or in evolutionary studies (Nei, 1987). Distance measures based on the SMM utilise size differences between specific alleles, and include Delta mu squared and the Pairwise squared distance (Goldstein *et al.* 1995). These distance measures are useful in evolutionary genetic distances, as they maintain a linear relationship to time over a long period, for example, several thousand generations. The linear relationship maintained by these distance measurements makes it an accurate means by which to study relationships over several thousand generations. For example, as previously noted in the Tang Annals, Islam was introduced into China in 651 AD (Wong and Dajani, 1988). Using the three Muslim populations under investigation and assuming a human generation span of 25 years, this equates to approximately 55 generations. It follows that the historical origins of these populations could be identified and correlated with the establishment of Islam into China.

However, the examination of present-day population divisions requires distance measures such as D_A (Nei, 1987), D_{ps} (Bowcock *et al.*, 1994), and D_S (Nei, 1972). These measures are based either on the frequencies of all alleles at a shared locus 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 of time, but are simply a measure of the current genetic relationship between populations. Similarly, the chord distance measure by Cavalli-Sforza uses a mathematical formula based on Pythagoras theory to calculate genetic diversity (Hartl and Clarke, 1997).

3.2.5 Diversity

A theoretical concept to accurately compare Y-chromosome and autosomal genetic diversity in the Bo'an, Salar and Dongxiang can be used. Pèrèz-Lezuan *et al.* (1997) proposed a direct comparison of Y-Chromosome and autosomal gene diversity using the following formula:

$$\mathbf{D_{au}} = \frac{\mathbf{4D_Y}}{\mathbf{(3D_Y - 1)}}$$

Where $\mathbf{D_{au}}$ is the autosomal gene diversity and $\mathbf{D_Y}$ is Y-Chromosome diversity. This method assumes that the effective number of Y-Chromosomes is one quarter the number of autosomal chromosomes. Gene diversity is mathematically analogous to expected heterozygosity but as the Y-chromosome is effectively haploid, this value represents the probability that two randomly selected alleles are from the same population.

Only a small portion of the Y-chromosome is available that can undergo limited recombination with the X-chromosome. An absence of recombination means genetic diversity is more limited in Y-chromosomes than in autosomal chromosomes. However, the lack of recombination increases the effect of genetic drift. It is this property of the Y-chromosome that could prove very useful in determining genetic differences.

Y-Chromosome microsatellite markers have proven useful in population genetics because of the ability to perform accurate haplotype analysis. A haplotype refers to a unique combination of genetic markers present in a chromosome (maternal or paternal) (Strachan and Read, 1999). Associations among the three study groups will be made on the basis of shared haplotype distributions.

3.2.6 Phylogenetic trees

To infer relationships between populations, distance measures are converted into phylogenetic trees. By this means sequence data from a population can be used to determine historical relationships with other populations. There are two types of phylogenetic tree: the population tree, which dates population split by measuring allele frequencies or distances, and the gene tree, which dates mutational origin by means of positive and negative detection. The gene tree offers greater genetic depth than the population tree (Hartl and Clark, 1997).

The techniques used to construct phylogenetic trees are based on algorithms that assume populations evolve in a predictable fashion, thus permitting estimation of the divergence between populations. A typical method of this type is the unweighted pair-group method with arithmetic mean (UPGMA). Using a pairwise distance matrix, the UPGMA technique assumes that all sequences evolve at the same rate. By grouping the smallest population with the smallest distance, and comparing mean distances between populations, a distance matrix is constructed. This mathematical process of comparison and grouping repeats itself until all populations are clustered into a tree with differing branch lengths (Hartl and Clark, 1997). Maximum parsimony can also be employed in the construction of phylogenic trees and is based on the effects of mutation. Using the smallest number of mutational events to account for evolution, maximum parsimony draws on sequences from a common ancestor to construct a tree. A third algorithm is used in the neighbour-joining method, which assumes that all sequences are related to one another. The neighbour-joining method is useful when the substitution rates in a sequence are varied or unknown (Hartl and Clark, 1997).

Bootstrapping is a common test for assessing the level of statistical confidence in each node of a phylogenetic tree. For each node of the tree the original data are re-sampled, a tree is drawn and the nodes are usually tallied up to a minimum of 1000 repetitions. The results are displayed as a number next to each node, indicating the percentage of occasions on which the cluster is present among the re-sampled trees. A high percentage value indicates high confidence in that node of the tree (Hartl and Clark, 1997).

3.2.7 Hardy-Weinberg equilibrium

The Hardy-Weinberg (HW) law is one of the fundamental concepts in population genetics, independently developed by the British mathematician Godfrey Hardy and the German physician Wilhelm Weinberg. HW states that, given a set of assumptions, allele frequencies will remain constant from generation to generation. These assumptions are:

1. Mating within the population occurs at random.

2. The population is infinitely large, which in practical terms means that the population is sufficiently large to ensure sampling errors and random effects are negligible.
3. There is no selective advantage for any genotype that is all genotypes produced by random mating are equally viable and fertile.
4. There is an absence of other factors, including mutation, migration, and random genetic drift.

If no mechanisms that can cause an evolutionary change are acting on a population, the gene pool frequencies will remain unchanged. Essentially, the Hardy-Weinberg law exists only in an ideal population and acts as a point of reference.

An equation was developed to represent the Hardy-Weinberg law and is referred to as the Hardy-Weinberg equilibrium (HWE). The HWE determines the genotype frequencies in a population and tracks their changes from one generation to another and is defined as follows:

$$\mathbf{p^2 + 2pq + q^2 = 1}$$

The HWE equation is defined in terms of a biallelic locus, where p is the frequency of the first allele and q the second allele for a locus consisting of a pair of alleles (A and a) (Hartl and Clark, 1997).

3.2.8 *Population inbreeding*

Inbreeding within a subpopulation is caused by non-random mating between individuals and often occurs between biologically related individuals. Two individuals are said to be related if among the ancestors of the first individual are one or more ancestors of the second individual. Related individuals will contain a large proportion of shared genes due to common descent, their offspring will display a higher level of homozygosity, and conversely, a lower level of heterozygosity than expected. Identical gene copies that are shared due to ancestry are known as being identical by descent (IBD). In assessing the degree of inbreeding within and between populations, the most common measure is the *F*-statistic, where *F* represents the probability that the offspring is homozygous due to IBD at a randomly chosen autosomal locus.

Wrights (1921) within subpopulations *F*-statistic can be used to estimate a ratio of the observed to expected heterozygosity where:

$$\boldsymbol{Fis} = \frac{\mathbf{H_E - H_O}}{\mathbf{H_E}}$$

Fis is the inbreeding coefficient, $\mathbf{H_E}$ is the average expected heterozygosity and $\mathbf{H_O}$ is the average observed heterozygosity estimated from each subpopulation (Hartl and Clark, 1997).

Population substructure can lead to inbreeding-like effects, such as a reduction in observed heterozygosity when compared to the expected heterozygosity. This effect is known as Wahlund's effect. This relationship shows that as allele frequencies in two subpopulations deviate, the average expected heterozygosity in those populations will always be less than that expected from the pooled allele frequencies ($\mathbf{H_T}$). A between subpopulations *F*-statistic can be estimated from this ratio:

$$\boldsymbol{Fst} = \frac{\mathbf{H_T - H_S}}{\mathbf{H_T}}$$

As allele frequencies deviate, the difference in $\mathbf{H_T}$ and $\mathbf{H_S}$ will increase and ***Fst*** will therefore also serve as a measure of genetic distance among subpopulations (Hartl and Clark, 1997).

$$\boldsymbol{Fit = \frac{H_T - H_I}{H_T}}$$

The inbreeding coefficient *Fit* is a measure of the correlation of alleles for the entire population and thus a combination of both the within and between subpopulation effects.

3.3 Method

3.3.1 Extraction of DNA

For each individual DNA was isolated from two blood spots using proteinase K treatment, followed by phenol/chloroform extraction and isopropanol precipitation at minus 20^{0}C overnight. The blood spots were cut from the filter paper, quartered, and placed in a 1.5ml microtube with 250 µl 0.1% Triton-100 and 15 µl 20mg/ml proteinase K at room temperature. The sample was mixed gently for 5 minutes before incubation at 50^{0}C for 30 minutes. Following incubation the sample was again gently mixed for 5 minutes, then vortexed for 1 minute before being incubated at 50^{0}C for an additional 30 minutes. On completion of the second incubation period, 25 µl of 10 x SET Buffer mix (500mM Tris pH 8, 50mM EDTA, 5% SDS) was added. The sample was mixed thoroughly and then incubated at 50^{0}C for 30 minutes. 500 µl of 1:1 chloroform/phenol was added and mixed by inversion for 1 minute. The sample was then centrifuged for 10 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 100% isopropyl alcohol were added to the supernatant. The tube was mixed and left overnight at –20^{0}C to precipitate the DNA from the solution. On the following day the sample tube was centrifuged at room temperature for 15 minutes at 13,000 rpm, the supernatant carefully discarded and the DNA pellet retained. 500 µl of ice-cold 70% ethanol was used to wash the DNA pellet. After inverting the microtube twice the sample was centrifuged for 10 minutes at 13,000 rpm. The ethanol was carefully discarded and allowed to evaporate at room temperature for approximately 1 hour. Finally, the pellet was resuspended in 50 µl of autoclaved distilled water.

3.3.2 Quantifying the DNA concentration of samples

The DNA sample concentration and purity was analysed using a Beckman™ DU 640 spectrophotometer, calibrated against distilled water. For each sample the optical density was measured using a 1:20 dilution of the DNA extract at wavelengths of 260 nm and 280 nm. After each sample was analysed, the cuvette was washed once with 70% ethanol and once with purified water (approximately 100 µl dH_20).

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. Quantification of the sample DNA (µg/mL) was obtained using the simplified formula:

$$\textbf{Total DNA (ng/mL)} = \textbf{OD}_{\textbf{260}} \textbf{ x 50 x dilution}$$

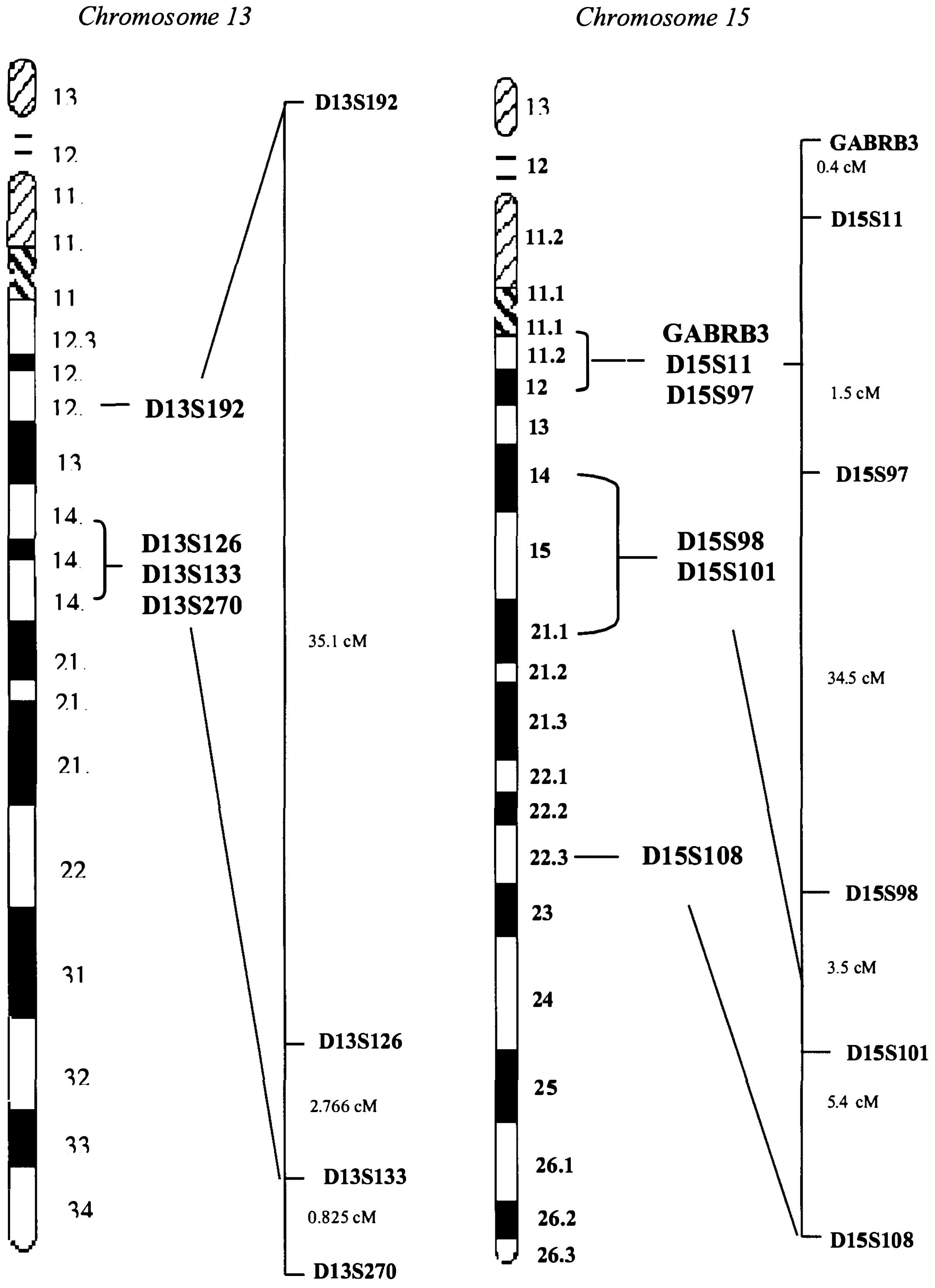
The spectrophotometric reading of total DNA (µg/mL) was used to produce a standardised dilution of 20 ng/µL DNA in any given volume of distilled water for each sample.

3.3.3 Microsatellite markers

The autosomal markers analysed in this study were chosen from a panel of markers originally recommended by Stanford University (see Appendix A). Figure 3.1

illustrates linkage and cytogenetic maps of chromosomes 13 and 15, indicating the positions of the markers used in this study. Y-chromosome markers used in the study were chosen from a panel of markers recommended by the Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University (http://ruly70.medfac.leidenuniv.nl/). The approximate locations of Y-chromosome markers used in the study are illustrated in figure 3.2. An exact position for the Y-chromosome markers under investigation has yet to be determined.

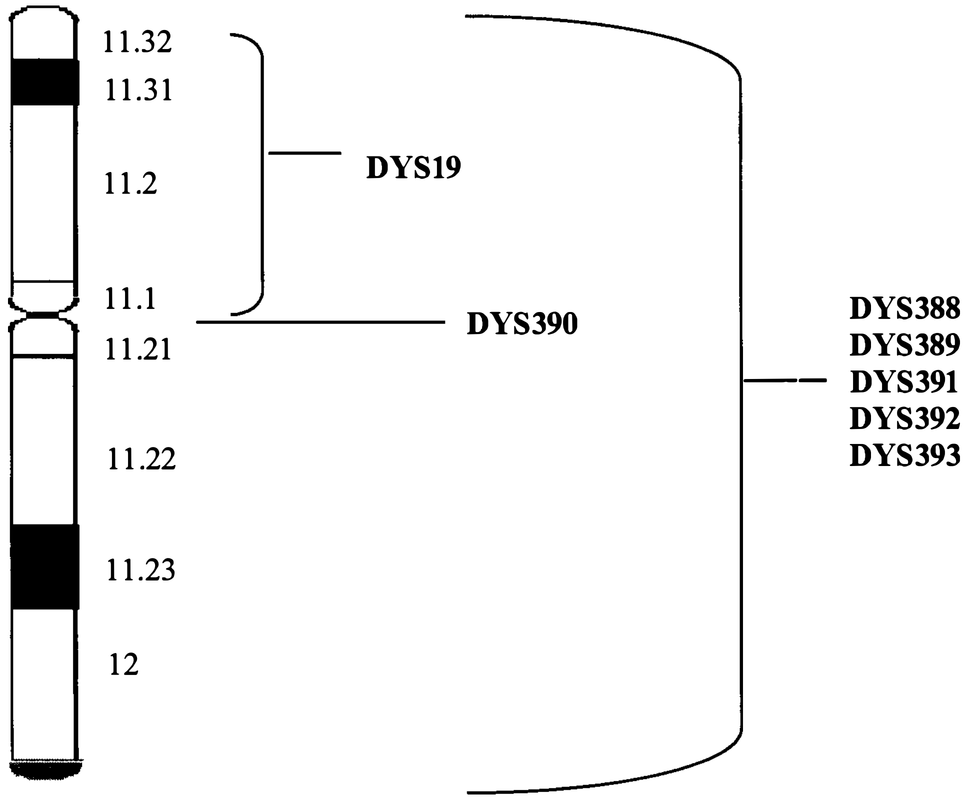
Figure 3.1 Comparative linkage maps locating microsatellite markers analysed from chromosomes 13 and 15. The distances are measured in centimorgans.



Chromosome 13

Chromosome 15

Figure 3.2 Cytogenetic map of the Y-chromosome indicating the approximate positions of the microsatellite markers used.



3.3.4 Polymerase chain reaction

The polymerase chain reaction was first reported some 30 years ago (Kleppe *et al.* 1971). Since then it has driven biological research given the capacity to quickly and accurately replicate specific regions of the human genome. The commercial development of heat stable polymerase (*Taq* polymerase) and thermo cycling instruments which can heat and cool samples repeatedly in a cyclical manner, has made PCR a very widely used technique.

The PCR reaction consists of three steps: denaturation, annealing and elongation. Heating at 95^0C denatures the template and disassociates the DNA from double to single strands. The mixture is then cooled to an optimal temperature to allow the oligonucleotide primers to bind to their complementary single stranded target sequences on the DNA template. Unidirectional DNA amplification is achieved by the extension of microsatellite primers with a supply of deoxynucleotide triphosphates (dNTPs) and a heat stable DNA enzyme such as *Taq* (*Thermus aquaticus*) polymerase. *Taq* polymerase can extend primers at temperatures of up to 72^0C, using the energy in the triphosphate bond to catalyse a reaction and allowing the formation of complementary bases effectively creating second strand. With each repeat cycle of denaturation, primer annealing and DNA synthesis, the number of copies of sequence between the forward and reverse primer sites is doubled, thus exponentially amplifying the region of interest.

3.3.5 PCR protocol

Autosome: A 10 μl reaction was prepared for each sample of genomic DNA. This consisted of 3 μl (20 ng/ μl) of target DNA, 1 μl of forward primer and 1 μl of reverse primer, 1 μl of 10 x buffer (containing 1.5mM of $MgCl_2$ solution [Perkin Elmer®]), 0.2 μl $MgCl_2$ (25 mM), 0.4 μl dNTPs, 0.1 μl of Amplitaq *Taq* polymerase (Perkin Elmer®) and 3.3 μl of dH_2O. Due to the variability and the quality of the target DNA and the fidelity of the primers, $MgCl_2$ concentrations varied between 1 mM and 4 mM. As detailed in appendix 4.1, the ten autosome markers were successfully amplified using the following touchdown conditions. The samples were initially denatured for 5 minutes at 94^0C. This was followed by 15 cycles of denaturing at 94^0C for 20 seconds, one minute of annealing starting at 63^0C and reducing in each cycle by 0.5^0C (giving a final temperature of 55.5^0C), and a 30 second extension period at 72^0C. A further 20 cycles followed, each consisting of 20 seconds denaturing at 94^0C, 20 seconds annealing at 55^0C, and 30 seconds of extension at 72^0C. The cycle concluded with a five-minute extension period at 72^0C.

Markers D15S98 and D15S97 were found to require lower annealing temperatures, and so a composite of the protocol given above was used with temperatures of $58\text{-}50^0$C for the touchdown phase and 50^0C for the extension phase.

Y-chromosome: A 10 μl reaction was prepared for each sample of DNA. This consisted of 5 μl (20 ng/μl) of target DNA solution, 1 μl of 10 x buffer solution (containing 1.5 mM of $MgCl_2$ [Perkin Elmer®]), 1 μl of $MgCl_2$ (25mM), 0.4 μl of 5mM dNTPs, 0.6 μl of forward and reverse primer containing 100ng/ μl of primer oligonucleotides, 0.1 μl of Hot Star™ *Taq* polymerase (Qiagen®) and 1.9 μl of dH_2O.

Qiagen® Hot Star™ *Taq* polymerase was used for Y-chromosome markers, as pre-experimentation with this enzyme produced greater amplification and fidelity of markers, most of which proved difficult to amplify.

Mitochondrial DNA: A 100 µl reaction was prepared for each randomly selected sample of DNA from the Sala, Bo'an and Dongxiang. This consisted of 5 µl (5–10 ng/µl) of target DNA solution, 16 µl of 10 x buffer solution (containing 1.5 mM of $MgCl_2$ [Perkin Elmer®]), 6 µl of $MgCl_2$ *(25mM),* 2.5 µl of 5mM dNTPs, 2 µl of forward and reverse primer containing 100ng/µl of primer oligonucleotides, 0.5 µl of Hot Star™ *Taq* polymerase (Qiagen®) and 77 µl of dH_2O. The following PCR conditions were used for 30 cycles: 94^{0}C for 45 seconds, 66^{0}C for 60 seconds and 72^{0}C for 60 seconds.

The PCR primers were composite oligonucleotides consisting of phage M13 and human mtDNA-specific sequences for the D-loop region of HV-1. Primers for HV-1 segment are shown in table 3.2, with the M13 sequence shown in lower case which is designed to assist the following sequence analysis (McBride *et al.*, 1989). Segment I sequence spanned nucleotide position 15997 to 16400 (Anderson *et al.,* 1981), with a maximum length of nucleotides 405 (allowing for insertions).

Table 3.2 M13 mtDNA primer sequences

HV-1 primer
5'-tgtaaaacgacggccagtTGATTTCACGGAGGATGGTG-3'
5'-caggaaacagctatgaccCTCCACCATTAGCATACCGCCA-3'

3.3.6 Agarose gel electrophoresis

Agarose gel electrophoresis is a routine technique in molecular biology for the analysis of DNA. Its many applications include the analysis of enzymatic manipulations such as restriction digestion, or ensuring that the microsatellite sequences used in a study are successfully amplified via PCR.

Agarose is a linear polymer extracted from seaweed. The preparation of agarose gels involves melting agarose powder in a 1 x TAE (0.04 M Tris-acetate, 0.001 M EDTA), an electrolytic buffer, until it becomes a transparent solution. This was achieved by heating the medium in a microwave oven at a high setting for approximately 1 minute. To prevent evaporation of water and a resultant increase in agarose concentration, care was taken not to boil the solution for long periods. The amount of agarose used was dependent on the fragments to be analysed or separated. Low percentage agarose gels (i.e. 1%) were used for the analysis of large fragments whereas higher percentages (i.e. 3%) were prepared for the separation and analysis of smaller fragments.

Once the agarose had completely dissolved in TAE it was allowed to cool to 60^0C in a hot box or a water bath, and 2 µl of ethidium bromide per 100 ml of agarose was then added to the solution. The solution was then poured into a Perspex gel tray taped at the open ends. Approximately 50 ml was poured into a small tray (8.5 cm x 11 cm) or 100 ml into a large tray (13.5 cm x 15 cm). A well-forming comb was placed at one end and positioned about 2 mm above the bottom of the tray. The gel was allowed to harden at room temperature for a minimum of 20 minutes. When set the agarose formed a matrix, the density of which was determined by the concentration of agarose.

The comb was removed and the gel slab placed in a horizontal electrophoresis tank and submerged in buffer (1 x TAE). DNA to be electrophoresed was combined with loading buffer, pipetted into the wells and an electric charge applied across the gel. The DNA, which is negatively charged due to its phosphate backbone, migrated towards the positive pole (anode).

The rate of migration of DNA through agarose gels is determined by three parameters. The first of these parameters is, the molecular size of the DNA. Linear, double stranded DNA molecules migrate through the agarose gel matrix at a rate that is inversely proportional to the $\log_{10}$ of the number of base pairs. Larger molecules have a slower rate of migration due to their greater frictional drag, and because they pass through the pores of the gel less efficiently than smaller molecules. Secondly, higher concentrations of agarose give a better separation of small DNA fragments. Finally, super helical, circular, nicked circular and linear DNA of the same molecular weight migrate through agarose gels at different rates (Maniatis *et al.*, 1982).

3.3.7 Agarose gel protocol

A standard 3% agarose gel was used in the analysis of autosomal, Y-chromosome and mtDNA. A 3% agarose gel comprising 3 g of agarose powder (Sigma Chemical Company) in 100 ml 1 x TAE buffer was prepared and poured into a gel tray.

Based on the quantity of DNA expected following amplification (e.g. autosomal product generally yielded more band illuminance in an agarose gel at smaller quantities than the Y-chromosome), 3-5 µl of the PCR products were loaded into each well with 1 µl of loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll [Type 400 Pharmacia]). 1.5 µl of *pUC19* DNA/HPA II (0.5 mg/ml; Biotech®; fragment size range 26 – 501 bp) size standard was loaded into lane one. The gel was electrophoresed at 80 V for approximately 20 minutes.

3.3.8 Visualisation of gel

Visualisation of DNA electrophoresed in an agarose gel was achieved by ethidium bromide (EB) staining and a Hoefer® Mighty Bright™ UV transilluminator. EB contains a planar group that intercalates between the stacked bases of DNA. The fixed position of this group and its close proximity to the bases causes dye bound to DNA to display an increased fluorescent yield compared to that of the dye remaining free in the solution. Ultraviolet radiation at 254 nm is absorbed by the DNA and transmitted to the dye while radiation at 302 nm and 366 nm is absorbed by the bound dye itself. In both cases the energy is re-emitted at 590 nm in the red-orange region of the visible spectrum. Since the fluorescent yield of DNA complexed with ethidium bromide is much greater than that of unbound dye, small amounts of DNA can be detected in the presence of EB in the gel (Maniatis *et al*., 1982).

3.3.9 Digital photography of agarose gels

A permanent record of the DNA visualised within the agarose gel was obtained by placing the EB stained gel on a UV transilluminator and photographing the image using the Kodak® DC120 Electrophoresis Documentation and Analysis System™. The resulting digital image could then be stored on a diskette or printed.

3.3.10 Genotyping

An ABI 373 DNA Sequencer™ with Genescan™ allele-sizing software was used to accurately genotype microsatellite alleles. The allele sizing technique is based on the fluorescently labelled primers detailed in table 3.3 incorporated into the microsatellite markers.

Table 3.3 Fluorescent molecules can be chemically bonded to microsatellites, 22-mer oligonucleotide forward primers were used.

Fluorescent label	**Colour**	**Molecular structure**
TET	Green	4, 7, 2, 7–tetrachloro-6-carboxyfluorescein
HEX	Yellow	4, 7, 2, 4, 5, 7–hexachloro-6carboxyfluorescein
FAM	Blue	6-carboxyfluorescein

For loading into a polyacrylamide gel, 0.5 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 standards are supplied as part of the Genescan-500 kit™ and the size standard is the result of a digestion of plasmid DNA with the restriction enzymes *Pst1* and *BstU1*. The subsequent DNA fragments are then labelled via chemical bonding with Tamra (N, N, N', N'-tetramethyl-6-carboxyrhodiamine).

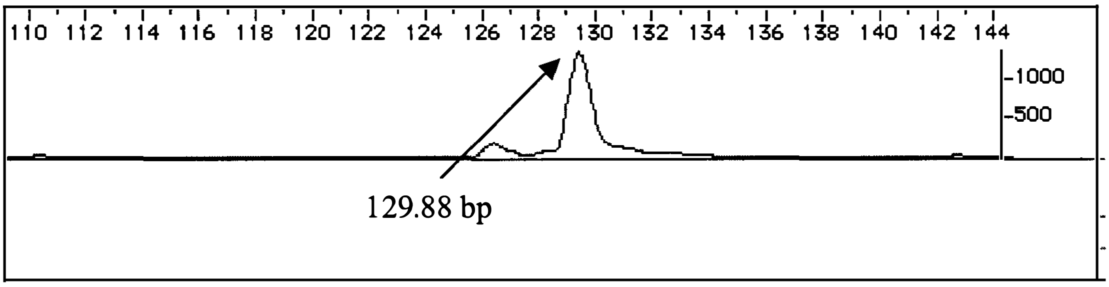
3.3.11 Polyacrylamide gel

Polyacrylamide is composed of the monomer acrylamide, which in conjunction ammonium persulphate, TEMED (N,N,N',N'-tetramethylethylene-diamine), and bisacrylamide (N.N'-methylenebisacrylamide) initiate a chain reaction that polymerises to form a gel. The length of the chains is determined by the concentration of acrylamide in the polymerisation reaction. In the presence of urea or formamide, base pairing of nucleic acids is suppressed allowing DNA to migrate through the gel at a rate that is independent of base composition and sequence. This variety of gel was chosen for the present study because of its superior resolving power than agarose gel; it can separate molecules of DNA whose lengths differ by as little as 0.2% (i.e. 1bp in 500bp) (Maniatis *et al.*, 1982).

The polyacrylamide gel comprise 40% acrylamide/bisacrylamide at a ratio of 19:1, 420.5 g of urea, and 100 ml of 10 x TBE buffer (890 mM Tris-borate, 2 mM NaEDTA, dH_2O), was made up to 1 litre with distilled water and stored at 4^0C. A 50 ml preparation of this gel was then mixed with 250 µl of 15% ammonium persulphate and 28 µl of TEMED. The mix was dispersed between two glass plates using a 100 ml syringe. The glass plates were separated by spacers at a thickness of 0.3 mm on either side and fixed together with bulldog clips with a well-to-read distance of 24 cm. A 50 well square toothcomb was positioned into the top of the gel apparatus, and the gel allowed to set for approximately 2 hours at room temperature. The gel plates were checked using the "plate check" setting in Genescan™ to identify any possible gel irregularities prior to loading. The gel was then pre-run for 10 minutes at 28^0C to optimise the temperature of the gel.

Following the pre-run, 2 µl of each sample solution were pipetted into the gel wells and the gel was run for 12 hours using filter setting B. The use of filter B allowed the marker colours, as identified in table 3.2, to be displayed. The resulting gel pictures were analysed and extracted via the Genescan™ software program, and the Genotyper™ software program used to analyse these data and to assign peaks to microsatellite alleles. As illustrated in figure 3.3, the results were presented as a series of peaks, with each microsatellite marker represented as one peak for homozygous and two peaks for heterozygous alleles. The allele sizes were recorded using Microsoft Excel™ version 7.

Figure 3.3 Genescan ™ output - homozygote peak



3.3.12 Sequencing

DNA sequencing was based on the dideoxynucleotide chain-termination method developed by Sanger *et al.* (1977). In this method, enzymatic sequencing relies on the ability of a thermostable DNA polymerase (i.e. *Taq polymerase*) to extend a primer hybridised to the template to be sequenced, until a chain-terminating nucleotide is incorporated. Four sequence reactions are determined, each of which contain a combination of deoxynucleotide triphosphate (dNTP) and dideoxynucleotide triphosphate (ddNTP). The ddNTPs lack the 3′-OH necessary for chain extension, thereby terminating the growing oligonucleotide selectively at A, G, C or T, depending on the presence of a dideoxy encountered in each reaction.

Each mtDNA PCR product (described previously) was purified to ensure that fragments were free of residue primers, nucleotides, polymerases, and salts. The Qiagen™ QIAquick™ PCR purification kit was used because it allowed DNA fragments ranging from 100bp to 10kb 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™ 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 into the collection tube discarded. The product that remained in the column (collection tube) was washed with 750 µl of PE buffer by centrifuge for 1 minute at 13,000 rpm. The flow through was then discarded and again centrifuged at 13,000 rpm for 1 minute. The flow through material and the collection tube were discarded and replaced with a new collection tube. The DNA was eluted using 20 µl of dH_2O, carefully placed at the centre of the columnar membrane and allowed to stand for approximately 1 minute, then centrifuged at 13,000 rpm for approximately 1 minute.

The PCR product was then sequenced using Applied Biosystems Ready Reaction Dye Primer™ (-21M13 and M13Rev) cycle sequencing kits. As shown in table 3.4, four single pre-mix cycle sequencing reactions were needed per sample, i.e. one for each of the bases A, C, G, and T. Bases A and C each consisted of 4 µl of premix, with 1 µl of eluted DNA added to a total volume of 5 µl. Bases G and T each consisted of 8 µl of premix with 2 µl of eluted DNA added to a total volume of 10 µl. The following oligonucleotides served as sequencing primers:

M13F:	5′-TGTAAAACGACGGCCAGT-3′
M13R:	5′-CAGGAAACAGCTATGACC-3′

Table 3.4 Components of the ABI PRISM™ Dye Primer Cycle Sequencing Ready Reaction Kit: pre-mix sequencing reactions.

Base	**Dye Primer** (-21M13 or M13Rev)	**Components**
Adenine (A)	JOE (green)	ddATP, dATP, dCTP, 7-deaza-2′-deoxyguanosine-5′-triphosphate, dTTP, Tris-HCl (pH 9.0 at 25^{0}C), $MgCl_2$, thermostable pyrophosphatase, Amplitaq DNA, FS.
Cytosine (C)	FAM (blue)	ddCTP, dATP, dCTP, 7-deaza-2′-deoxyguanosine-5′-triphosphate, dTTP, Tris-HCl (pH 9.0 at 25^{0}C), $MgCl_2$, thermostable pyrophosphatase, Amplitaq DNA, FS.
Guanine (G)	TAMRA (yellow)	ddGTP, dATP, dCTP, 7-deaza-2′-deoxyguanosine-5′-triphosphate, dTTP, Tris-HCl (pH 9.0 at 25^{0}C), $MgCl_2$, thermostable pyrophosphatase, Amplitaq DNA, FS.
Thymine (T)	ROX (red)	ddTTP, dATP, dCTP, 7-deaza-2′-deoxyguanosine-5′-triphosphate, dTTP, Tris-HCl (pH 9.0 at 25^{0}C), $MgCl_2$, thermostable pyrophosphatase, Amplitaq DNA, FS.

The sequencing reactions were placed in a Perkin Elmer 9600™ thermal cycler utilising the following conditions: 15 cycles of denaturing at 96^0C for 10 seconds, 5 seconds of annealing at 55^0C, and a 1 minute extension period at 70^0C. A further 15 cycles followed, each consisting of 10 seconds denaturing at 96^0C and 1 minute of extension at 70^0C. The cycle concluded with an infinite 4^0C hold until collection.

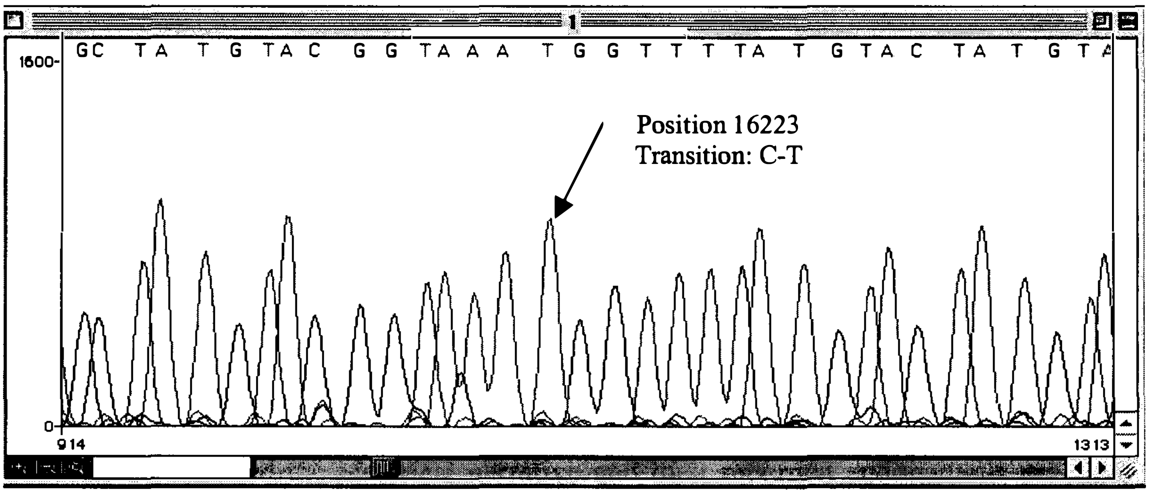
Following cycle sequencing, the DNA sample extension product was precipitated and concentrated in an autoclaved microtube, with 80 µl of 95% ethanol added. The four extension reactions for each sample (i.e. A, C, G, and T) were transferred into one tube and then mixed with the ethanol. The mixed extension products and ethanol were then placed on ice for approximately 15 minutes to facilitate precipitation. The samples were next centrifuged at 13,000 rpm for approximately 30 minutes. The supernatant was then removed and 250 µl of 70% ethanol added to rinse and de-salt the sample. The resultant sample was centrifuged at 13,000 rpm for 5 minutes. Following this procedure the supernatant was discarded and the DNA pellet was left at room temperature until completely dry.

The pellet sample was next prepared for loading into the polyacrylamide apparatus. Loading buffer (25 mM EDTA [pH 8.0], blue dextran 50 mg/ml) and deionised formamide was mixed to a volume of 5:1 respectively, and aliquots of 3 µl were combined with each pellet sample. The polyacrylamide gel components were described previously with the Genescan™ methodology. In preparing a sequencing polyacrylamide gel, the glass plates are required to have a well-to-read distance of 48 cm and a 48 well sharp toothcomb is placed in the top of the gel apparatus. The gel plates were checked using the "plate check" setting in Sequencescan™ to identify any possible gel irregularities prior to loading, and pre-run for 10 minutes to optimise the gel temperature. Following the pre-run, 2 µl of each sample solution was pipetted into the gel wells and the gel was run for hours using filter setting A. The use of filter A allowed the marker colours, as identified in table 3.3, to be displayed. The resulting gel pictures were analysed and extracted with the Sequencescan™ software program.

Once the gel pictures had been satisfactorily extracted, further analysis could be undertaken using the Sequence Navigator version 1.01 (ABI Prism™) software program. For each individual sample firstly the forward M13 Primer sequence was selected, and then the complementary reverse M13 sequence were read in a parallel 5′-3′ direction. The two sequences were comparatively aligned and identified by locating the reverse primer sequence (5′-3′) and the forward primer sequence (3′-5′), which effectively identified the common sequence between the forward and reverse primers. All nucleotide bases at either side that were not considered a part of the sequence under investigation were discarded. The sequences were again checked to ensure accurate alignment of complementary bases.

As illustrated in figure 3.4, the height of each peak was used to identify the relative intensities of each of the four fluorescently labelled ddNTPs (see table 3.3). Where fluorescent peaks stood independently, with anomalous factors such as signal irregularities and background noise at a minimum, the nucleotide base was taken as being correct. If the nucleotide base was identified as ambiguous and not accurately representing the sequence (i.e. low signal strength), an ‘N’ was used to identify the base position as unresolved. Once the position had been resolved the ‘N’ was replaced by a lower-case letter (i.e. a, c, g, and t). Resolution of a nucleotide base was achieved by consultation with the complementary sequence.

Figure 3.4 Sequence navigator software: mtDNA identification of a polymorphism


1600-
GC TA T G TA C G G TA A A T G G T T T TA T G TA C TA T G T A
0
914
1313

Position 16223
Transition: C-T

Chapter 4

Results

4.1 Introduction

The objective of this section is to detail the results of the genome-based research undertaken on the Bo'an, Salar, and Dongxiang. These random sample populations were compared in terms of within- and between population genetic variation. The major aim was to gauge the degree to which the genetic structure of the three sample populations matched the historical narrative and their recorded endogamous practices. Therefore, special emphasis was placed on comparisons of homozygosity, allele distribution patterns, the effects of reproductive isolation, and migration, based on the autosomal, Y-chromosome and mitochondrial DNA gene pools.

4.2 Autosomal gene pool

The three populations were genotyped across ten autosomal loci. As indicated in table 4.1a, a total of 160 different alleles were identified among 212 individuals, with 53 alleles (33.1%) shared by all three communities. The reference population (CEPH or GDB) showed a similar mean number of alleles when compared with the three study populations. Distinctive allele distributions were recorded in the study populations with significant inter-community differences at two (D13S133, D13S270) of the ten autosomal loci (p<0.05). Loci D15S101 and D15S97 showed unimodal allele distributions, D13S133, D13S192, D13S270, D15S108 loci were bimodal, and D13S126, D15S11, D15S97, GABRB3 loci were multimodal. The 243 bp allele at locus D15S11 had the highest frequency (range, 0.557 – 0.661) across all populations. The number of alleles observed at each locus varied by STR marker, from a mean of 6.3 alleles for D13S126 to a mean of 18.0 alleles for D13S133.

Table 4.1a Comparison of allele numbers at autosomal loci in the Bo’an, Salar, and Dongxiang

Locus	No. of alleles			
	Bo'an	Salar	Dongxiang	CEPH/ GDB
D13S126	6	7	6	7
D13S133	21	14	12	15
D13S192	13	21	12	15
D13S270	9	12	9	6
D15S101	12	13	15	9
D15S108	13	10	14	11
D15S11	12	10	10	11
D15S97	10	11	9	10
D15S98	16	10	15	17
GABRB3	13	10	10	13
Mean	*12.5*	*11.8*	*11.2*	*11.4*

Table 4.1b Comparison of allele size at autosomal loci in the Bo'an, Salar, and Dongxiang

Locus	Allele size range			
	Bo'an	Salar	Dongxiang	CEPH/ GDB
D13S126	102-112	104-126	104-112	100-112
D13S133	127-189	126-176	124-176	131-187
D13S192	96-120	75-119	97-117	89-123
D13S270	69-95	77-95	73-91	81-95
D15S101	88-126	96-126	98-137	110-134
D15S108	131-163	139-165	137-165	141-161
D15S11	241-269	241-259	239-265	238-260
D15S97	170-188	171-187	171-187	168-186
D15S98	143-171	145-169	145-175	141-161
GABRB3	173-201	181-199	181-199	171-201

Table 4.1c Comparison of allele homozygosity at autosomal loci in the Bo'an, Salar, and Dongxiang

Locus	Homozygosity			
	Bo'an	Salar	Dongxiang	CEPH/ GDB
D13S126	0.299	0.231	0.274	0.33
D13S133	0.374	0.293	0.218	0.18
D13S192	0.184	0.214	0.129	0.10
D13S270	0.358	0.487	0.322	0.14
D15S101	0.299	0.237	0.143	0.16
D15S108	0.448	0.423	0.345	0.30
D15S11	0.358	0.474	0.468	0.33
D15S97	0.164	0.300	0.233	0.39
D15S98	0.246	0.313	0.259	0.08
GABRB3	0.522	0.176	0.278	0.27
Mean	*0.325*	*0.315*	*0.267*	*0.228*

The allele frequency distributions were similar in the Bo'an and Salar, but the Dongxiang showed a number of results that differed from the other two communities. Specifically, at seven loci there were two alleles in the Dongxiang that were higher in frequency than the corresponding alleles in the Bo'an and Salar (D13S192 alleles 97, 103; D15S101 alleles 100, 106; D15S108 alleles 145, 159; D15S11 alleles 243, 259; D15S97 alleles 179, 183; D1598 alleles 153, 157; GABRB3 185, 187).

The range of allele sizes, shown in table 4.1b, varied only slightly between the three populations. Loci could be differentiated by the mean allele size ranges, for example, loci D15S101 with a 9bp longer allele (CEPH/GDB, 24bp) and D15S11 with a 15bp shorter (CEPH/GDB, 22bp).

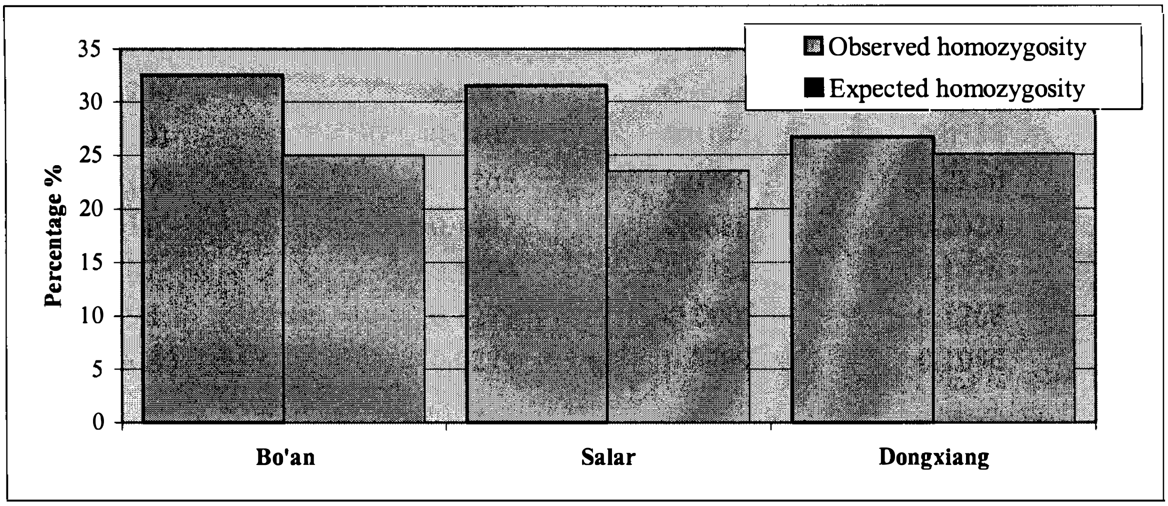
As shown in table 4.1c, an elevated pattern of homozygosity was evident in all three-study populations by comparison with the reference population. Mean observed homozygosity was 32.5% for the Bo'an, 31.5% for the Salar and 26.7% for the Dongxiang, compared with a value of 22.8% for the outbred reference population. For the Bo'an and Salar, at individual loci, except D15S97 and GABRB3, all showed substantially higher homozygosity values than the reference population.

Table 4.2 Estimation of exact Hardy-Weinberg P-values

Locus	Bo'an *p*-value	Salar *p*-value	Dongxiang *p*-value
D13S126	0.2648	0.0137	0.1068
D13S133	0.0000	0.0000	0.0000
D13S192	0.4086	0.0000	0.1734
D13S270	0.0020	0.0000	0.0038
D15S101	0.0075	0.0345	0.0000
D15S108	0.0000	0.0000	0.0090
D15S11	0.3583	0.0000	0.5310
D15S97	0.2952	0.0177	0.2905
D15S98	0.0372	0.0356	0.1813
GABRB3	0.0000	0.0701	0.9503

All three communities recorded significant deviations from Hardy-Weinberg equilibrium (HWE) at various loci. As illustrated in table 4.2, at the ten loci studied four (D13S133, D13S270, D15S101, D15S108) showed significant deviations from HWE across all three populations. In addition, significant deviations from HWE were observed at eight (p<0.05) of the remaining eighteen loci. Two of these loci were in the Bo'an, six in the Salar, and none in the Dongxiang (table 4.2).

Figure 4.1 Mean observed and expected number of homozygotes in the study populations



As shown in Figure 4.1, the average observed homozygosity levels in the Bo'an (32.5%) and Salar (31.5%) was greater than in the Dongxiang (26.7%). The average expected homozygosity levels in the Bo'an (25.02%) and Salar (23.58%) were similar to the result indicated for the Dongxiang (25.14%). Within the Bo'an and Salar the mean expected homozygosity values were lower than the observed homozygosity levels. The Dongxiang showed an expected homozygosity, across all loci in keeping with the observed homozygosity value.

An analysis of autosomal heterozygosity using the HWE test, under the hypothesis of a heterozygote deficiency, showed significant results at seven of the ten loci tested in the Bo'an, six of the ten loci tested in the Salar, and two of the ten loci tested in the Dongxiang. As expected, across all loci high deviations in HWE given heterozygote deficit were the opposite of the less deviation results recorded for HWE with heterozygote excess hypothesis (see Appendix).

Table 4.3 *F*-statistics for individual autosomal loci in the Bo'an, Salar, and Dongxiang

Locus	*Fis*	*Fst*	*Fit*
D13S126	0.0269	0.0735	0.0984
D13S133	0.1564	0.1060	0.2458
D13S192	0.0727	0.0747	0.1420
D13S270	0.1662	0.1668	0.3052
D15S101	0.0400	0.0144	0.0538
D15S108	0.2268	0.0740	0.2840
D15S11	0.1447	0.0680	0.2029
D15S97	0.1094	0.0881	0.1879
D15S98	0.1272	-0.0005	0.1268
GABRB3	0.0279	0.0799	0.1056
Mean	*0.1095*	*0.0848*	*0.1850*

As shown in table 4.3, the mean *Fis* values (Weir and Cockerham, 1984) calculated showed the higher inbreeding effects in the Bo'an (0.12) and Salar (0.16), whereas the Dongxiang (0.01) is the lowest among the three groups (see Appendix). These *Fis* values indicate that in the Bo'an and Salar inbreeding effects were predominantly within-community rather than between-community.

The mean percentage of private alleles, i.e. those unique to each community, was 9.61%. Table 4.4 identifies the private alleles per locus in all three populations. When each of the unique alleles were pooled together and compared with the total number of alleles per locus in all three communities, the values ranged from 9.09% at D15S97 to 38.09% at D13S192.

Table 4.4 Private autosomal alleles per locus in the study populations

Locus	Bo'an (n = 67)	Salar (n = 81)	Dongxiang (n = 64)	Total alleles per locus
D13S126	0	1	1	8
D13S133	5	0	6	25
D13S192	0	8	0	21
D13S270	1	2	1	14
D15S101	2	2	3	19
D15S108	1	1	2	16
D15S11	2	3	1	17
D15S97	0	1	0	11
D15S98	2	0	1	17
GABRB3	4	0	0	14

The measures of genome diversity observed in the three communities are presented in table 4.5, as (a) mean squared distance (d^2) and (b) relative mean squared distance (Rd^2). Based on the Stepwise Mutation Model (SMM), d^2 and Rd^2 are based on distances calculated between STR alleles within an individual (Shriver *et al.,* 1993). The lowest mean squared distances were calculated for the Dongxiang (d^2 = 61.66), by comparison with the Bo'an (d^2 = 178.63), and the Salar (d^2 = 110.45). When averaged across loci the mean squared distance may however produce misleading values, because of the disproportionate influence of certain loci on the average (Coltman *et al.,* 1998). For example, in the present study locus D13S133 contributed a much larger absolute d^2 value than the other loci examined (table 4.5a).

To overcome this problem, the mean squared distance at each locus was averaged across all three communities and its inverse value used as a locus-specific weight to calculate Rd^2. As shown in table 4.5b the Dongxiang again demonstrated the lowest average Rd^2 (11.131), by comparison with the Bo'an (Rd^2 = 27.339), and the Salar (Rd^2=14.426), emphasising the strong dynamic inbreeding process in Dongxiang that has not been provided by classical genetic parameter, homozygosity.

Table 4.5 Measures of genomic diversity based on (a) the mean squared distance and (b) the relative mean squared distance between STR alleles within an individual in the three communities.

(a) Mean squared distance

Locus	Bo'an	Salar	Dongxiang
D13S126	21.40	29.60	18.00
D13S133	949.80	454.50	19.40
D13S192	162.40	100.90	107.30
D13S270	87.90	41.70	60.70
D15S11	56.90	75.80	46.60
D15S97	116.80	141.60	104.60
D15S98	158.30	114.10	92.20
D15S101	44.20	48.90	30.20
D15D108	130.40	51.90	97.90
GABRB3	58.20	45.50	39.70
Average	*178.63*	*110.45*	*61.66*

(b) Relative mean squared distance

Locus	Bo'an	Salar	Dongxiang
D13S126	3.194	3.795	2.903
D13S133	141.761	58.269	3.129
D13S192	24.239	12.936	17.306
D13S270	13.358	5.346	10.194
D15S11	11.045	9.821	7.677
D15S97	17.433	18.795	16.871
D15S98	24.881	15.090	14.871
D15S101	8.194	6.654	5.613
D15D108	20.224	7.231	17.306
GABRB3	9.060	6.321	15.440
Average	*27.339*	*14.426*	*11.131*

4.3 Y-chromosome gene pool

The five Y-chromosome STRs present in the three communities are listed in table's 4.6a and b. A total of 23 different alleles were identified among 137 male individuals, with 14 alleles (60.9%) shared by all three communities. Distinctive allele distributions were recorded, with significant differences across four of the five Y-chromosome loci (p<0.02).

Table 4.6a Comparison of the number of alleles at Y-chromosome loci for the Bo'an, Salar and Dongxiang

Locus	No. of alleles			
	Bo'an	Salar	Dongxiang	L.U. forensic lab
DYS19	3	2	2	10
DYS388	4	6	6	7
DYS389	3	3	4	7
DYS391	2	4	3	6
DYS393	4	6	4	5
Mean	*3.2*	*4.2*	*3.8*	*7*

Table 4.6b Comparison of allele sizes at Y-chromosome loci for the Bo'an, Salar and Dongxiang

Locus	Allele size range			
	Bo'an	Salar	Dongxiang	L.U. forensic lab
DYS19	190-206	190-202	190-202	174-210
DYS388	122-134	122-137	122-137	125-143
DYS389	247-251	247-255	247-259	239-263
DYS391	283-287	279-291	279-287	275-295
DYS393	119-131	99-131	115-127	115-131

The patterns of allele distribution at all five loci (DYS19, DYS388, DYS389A, DYS391, DYS393) were unimodal. The most common allele was allele 190bp at locus DYS19 in the Dongxiang and Sala with a frequency of 0.885 and 0.778 respectively; and the Bo'an both alleles 128bp (DYS388) and 251bp (DYS389A) recorded a frequency of 0.700. The number of alleles observed at each locus varied by STR marker, from a mean of 2.3 alleles at DYS19 to a mean of 5.3 alleles at DYS388.

Table 4.7 Unique Y-chromosome alleles per locus

STR	DYS19	DYS388	DYS389A	DYS391	DYS393
Bo'an (n = 40)	1	0	0	0	0
Salar (n = 50)	0	0	0	1	1
Dongxiang (n = 47)	0	0	1	0	0
Total alleles per locus	*3*	*6*	*4*	*4*	*6*

The mean percentage of private alleles, i.e. those unique to each community was 17.3%, derived from table 4.7. Two previously unreported alleles were detected: an 122bp DYS388 allele in all three communities (frequencies= 0.700, Bo'an; Salar, 0.680; Dongxiang, 0.578) and a 99 bp DYS393 allele found only in the Salar at a frequency of 0.265.

The average gene diversity value was also calculated (Nei, 1987). For the Bo'an average gene diversity for the Y-chromosome loci was 0.4230, for the Salar 0.4610, and for the Dongxiang 0.4630 (see Appendix G).

As illustrated in table 4.8, the accumulated Y-chromosome *Fst* value for the Bo'an, Salar and Dongxiang is 0.3478, which indicates significant inter-population diversity.

Table 4.8 Accumulated *Fst* values for each locus in the Bo'an, Salar and Dongxiang

Locus	F*st*
DYS19	0.1114
DYS388	0.0156
DYS389	0.0572
DYS391	0.0004
DYS393	0.1632
All	0.3478

Pairwise *Fst* distances were calculated for all Y-chromosome loci between the three populations with values for DYS19 and DYS393 shown in table 4.9. The pairwise *Fst* measures nucleotide distances between the same loci among two groups and is a measure of inter-community diversity. A high figure would suggest greater diversity than a low figure. The lowest pairwise *Fst* distance value was obtained between the Dongxiang and Salar (0.0020) at locus DYS19, and the highest between the Salar and Bo'an (0.2333) at locus DYS393.

Table 4.9 Y-chromosome mean pairwise *Fst* distances for DYS19 and DYS393

DYS19	Bo'an	Salar
Salar	0.0830	
Dongxiang	0.2130	0.0020

DYS393	Bo'an	Salar
Salar	0.2333	
Dongxiang	0.0033	0.1977

As illustrated in table's 4.10a, b and c, among the three study populations, 50 Y-chromosome haplotypes were identified, with 35 (70%) haplotypes being community-specific. A total of 6 haplotypes were unique to the Bo'an, 13 to the Salar, and 16 to the Dongxiang. Unavailable haplotype data was recorded for several individual samples in the Bo'an (n = 1), Salar (n=7) and Dongxiang (n = 11), due to unsuccessful PCR amplification for the particular Y-chromosome STR loci.

Haplotype diversity could be measured using the mean pair-wise difference, which repeatedly calculates the distances between all the varied co-repeats across all the varied haplotypes. The mean pairwise difference values were 2.117 (SD ± 1.204) for the Bo'an, 2.303 (SD ± 1.283) for the Salar, and 2.317(SD ± 1.291) for the Dongxiang.

Table 4.10a Y-chromosome haplotype distributions in the Bo'an

Haplotype	Mean frequency	Distribution	Repeat per loci				
B 1	0.1026	4	14	12	13	11	13
B 2	0.0256	1	14	13	12	11	12
B 3	0.1538	6	18	12	13	10	12
B 4	0.0256	1	17	13	13	10	13
B 5	0.0513	2	17	12	13	10	12
B 6	0.1538	6	17	12	12	10	12
B 7	0.0769	3	17	13	13	11	13
B 8	0.0256	1	14	12	12	10	14
B 9	0.0769	3	17	12	13	11	13
B10	0.1538	6	14	12	12	10	12
B11	0.0513	2	17	12	13	10	13
B12	0.0513	2	14	14	13	10	13
B13	0.0513	2	14	14	13	10	12
	1.0000	39					

Table 4.10b Y-chromosome haplotype distributions in the Salar

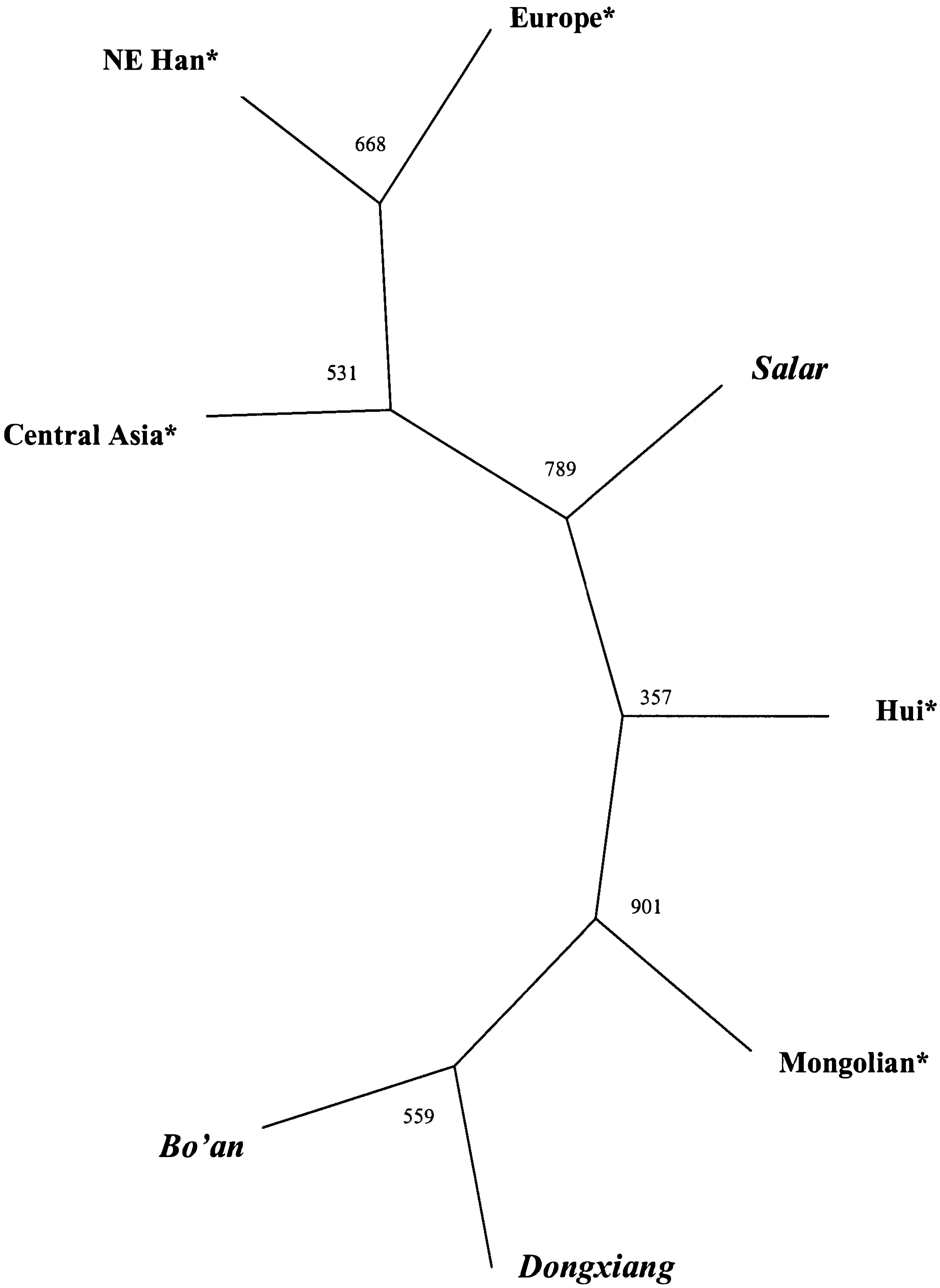
Haplotype	Mean frequency	Distribution	Repeat per loci				
S 1	0.0465	2	14	12	14	10	7
S 2	0.0233	1	17	13	13	10	11
S 3	0.0698	3	14	15	13	11	11
S 4	0.0233	1	14	12	12	12	11
S 5	0.0233	1	17	12	13	11	11
S 6	0.0698	3	17	12	12	10	11
S 7	0.1163	5	14	12	13	10	11
S 8	0.0465	2	17	12	12	10	12
S 9	0.0698	3	14	12	14	11	7
S10	0.0465	2	14	14	13	10	7
S11	0.0233	1	14	12	13	10	13
S12	0.0465	2	14	10	12	10	11
S13	0.0930	4	14	12	13	11	7
S14	0.0930	4	14	12	13	11	13
S15	0.0465	2	14	12	12	10	11
S16	0.0698	3	14	12	13	11	14
S17	0.0233	1	14	14	14	10	13
S18	0.0233	1	14	14	14	9	11
S19	0.0465	2	17	14	13	10	7
	1.0000	43					

Table 4.10c Y-chromosome haplotype distributions in the Dongxiang

Haplotype	Mean frequency	Distribution	Repeat per loci				
D 1	0.0556	2	14	14	13	10	12
D 2	0.0833	3	17	12	13	11	13
D 3	0.0278	1	14	12	13	11	11
D 4	0.0556	2	14	12	13	10	11
D 5	0.0278	1	14	13	13	10	13
D 6	0.0278	1	14	10	13	11	12
D 7	0.1389	5	14	12	12	11	13
D 8	0.0556	2	14	12	12	10	11
D 9	0.0278	1	14	15	13	10	12
D10	0.0556	2	14	13	13	11	13
D11	0.0278	1	14	12	14	10	12
D12	0.1389	5	14	12	12	10	13
D13	0.0278	1	14	13	14	11	14
D14	0.0278	1	14	12	15	9	13
D15	0.0278	1	14	13	14	10	13
D16	0.0556	2	14	12	12	11	14
D17	0.0556	2	14	15	12	10	12
D18	0.0833	3	14	10	12	10	12
	1.0000	36					

The phylogenetic relationships of populations can be constructed and represented as a dendogram or tree. Figure 4.2 illustrates an example of an unrooted Y-chromosome dendogram constructed to examine the relationship between the three study populations and European, Central Asian, Mongolian, NE Han, and Hui populations previously reported in the literature (de Knijff *et al.*, 1997). The tree typology is based on Nei's standard genetic distance, utilising the neighbour-joining method and 1000 bootstraps to assess the level of confidence in each node of the tree. The tree clearly places the Salar closer to Central Asian, European and Han Chinese populations, reflecting their Turkic-speaking origins. In contrast, the Bo'an and Dongxiang are both Mongolian-speaking communities and are closer to the Mongolian populations. The Hui form the largest component of the Muslim community in PR China and their intermediate positioning on the tree can be interpreted as reflecting a composite gene pool with diverse origins (Black *et al.*, 2000).

Figure 4.2 Y-chromosomal STR dendogram



* de Knijff *et al.* (1997)

4.4 Mitochondrial DNA (mtDNA)

A 360-nucleotide sequence in hypervariable region I (HV-I) of the mitochondrial D-loop region (position 16024-16383 in the reference sequence; Anderson *et al.,* 1981) was analysed in a total of thirty samples, comprising ten each from the Bo'an, Salar and Dongxiang communities. As illustrated in figure 4.3, among the three populations at least 2 sequences appeared constant, a sequence characterised by a T at position 16223 and a C at position 16362. In addition, 24 mtDNA sequences were unique, 19 mtDNA sequences were found once, two sequences were found twice (Bo'an 10 and 53, Salar 32 and 4) and one sequence was found three times (Salar 30, 31, 50). The total number of polymorphic sites identified within each population was for the Bo'an (n = 25), Salar (n = 17), and Dongxiang (n = 20). The total number of polymorphic sites identified among the populations was 44, or 12.2% of the overall 360bp reference sequence.

In position 16223, there was a C in the Anderson reference sequence and a high frequency of transition to T in the samples examined (Bo'an, 70%; Salar, 90%; Dongxiang, 70%). A high level of polymorphism also existed at position 16362, which presents as a T in the reference sequence but showed a high frequency of mutation to C in each of the study communities (Bo'an, 70%; Salar 60%; Dongxiang, 30%). Nucleotide diversity for the three populations was 0.0162, with a pairwise nucleotide difference for the mtDNA sequences of 5.90 (SD = 2.465).

Figure 4.3 Comparative alignment of mitochondrial DNA sequences

```
Bo'an  4     .G..T..A.GC......................C..........
Bo'an  6     ................AT...............C..........
Bo'an  7     ...................TT.....................C.
Bo'an 10     ...................T.......T.........A....C.
Bo'an 11     .......A...........T......................C.
Bo'an 14     ...................T..........T.............
Bo'an 19     .......A.....T.T...T.........T.....C.....C.C
Bo'an 32     ...C..C........................C......T.....
Bo'an 38     .......A......A....T.......T.........A......
Bo'an 53     ...................T.......T.........A....C.

Salar  2     ........A...T........C...........CG.........
Salar  4     ............T.....TT................G.....C.
Salar 13     ...C...A...........T...........C......T.....
Salar 14     ...................T........................
Salar 27     .....T.A...........T........................
Salar 29     ..........C........T......................C.
Salar 30     ...........T.......T...............C......C.
Salar 31     ...........T.......T...............C......C.
Salar 32     ............T.....TT................G.....C.
Salar 50     ...........T.......T...............C......C.

Dongxiang  7 .....T.A.................T.............C....
Dongxiang  8 ...................T...................C..C.
Dongxiang 10 ......C......T..................G...........
Dongxiang 13 .....T.A.................T.......C..........
Dongxiang 24 A.........C........T....................T.C.
Dongxiang 25 ...................T......T...............C.
Dongxiang 41 .....T.A...........T...AT.T.................
Dongxiang 45 ..C..T.A...........T..G.....................
Dongxiang 46 ...................T........A...............
Dongxiang 47 ...................T.........T..............

ANDERSON     GAATCCTGGATCCCGCACCCCTACCCACCCCTATATAGCTATTA

             11111111111111111111111111111111111111111111
             66666666666666666666666666666666666666666666
             00001111111111222222222222222222333333333333
             46790122467899111222344566899999000111244567
             26838169522423346123436716902458049169713724
```

mtDNA position

Chapter 5

Discussion

5.1 Introduction

The Peoples Republic of China is a culturally diverse nation composed of 56 separate ethnic groups, 10 of which are officially recognised Muslim minorities. The Bo'an, Salar, and Dongxiang populations are typical of many of the present-day Chinese Muslim communities. The paternal origins of these populations can be traced to Arab, Iranian and Central Asian traders of the Silk Road, and/or the Mongol peoples. Although, their individual histories and population sizes vary as previously indicated, it is believed that extensive inter-marriage with Han females occurred. Through time, each community retained many traditional customs, religious beliefs, and most notably, the language of their founders. The Salar are Turkic in origin and the Bo'an and Dongxiang are Mongolian-speaking. To identify the contributory founder populations and the underlying population dynamics, the present study was conducted into the genetic profile of each community, based on autosomal, Y-chromosome and mitochondrial DNA markers.

5.2 Autosome diversity

Autosomal analysis indicated elevated levels of homozygosity among all three communities by comparison with the reference populations (see table 4.1c). The Salar and Bo'an showed higher homozygosity levels suggestive of high levels of endogamous and/or consanguineous marriages. In contrast, the Dongxiang indicated lower values of homozygosity that indicated a lower level of endogamy. To fully understand the patterns of genetic diversity in the Bo'an, Salar, and Dongxiang communities, it is appropriate to note the variance in population size of the three communities (n = 11,683; n = 87,546; n = 373,669, respectively) (Family Planning Commission, 1998). Thus when collecting random samples, the probability of inadvertently selecting related individuals may be higher in the Bo'an and Salar than the Dongxiang. The combined effect of restricted local population sizes and endogamy could explain the elevated homozygosity values and the results obtained with HWE analysis.

Both the Bo'an and Salar shared significantly lower expected than observed homozygosity whereas in the Dongxiang samples these values were very similar. This would suggest that the Dongxiang may be more diverse in genetic terms than the relatively cohesive genetic structures of the Bo'an and Salar communities. A possible explanation could be that as the Dongxiang were the first of the three Muslim groups to establish themselves during the Sung Dynasty from 960-1270 AD, their longer timeframe as a community might have allowed greater opportunity for diversification of their gene pool. In addition, the Dongxiang are numerically the largest of the three populations, which could reflect both a large founding population and significant admixture through time.

The mean *Fis* probability values for the Bo'an and the Salar, also indicated higher intra-community values for consanguineous unions, in contrast to the lower *Fis* value exhibited by the Dongxiang. This effect in the Bo'an and Salar could probably be due to factors such as a limited choice or availability of a marriage partner or possibly as a result of social customs and religious practices in each community. While these factors would also be relevant to the Dongxiang, the concept of a larger and older population has permitted over time the migration of individuals to dilute down the Dongxiang gene pool. This is supported by the high homozygosity levels exhibited by the Bo’an and Salar and the lower levels exhibited by the Dongxiang. However, similar social and religious beliefs coupled with the co-residence of these three communities, would also infer some form of inter-community association. Clearly, the low autosome mean *Fst* value indicates a common autosomal structure among the three groups in this study, which would probably be as a result of inter-community marriages. As a consequence, the autosomal gene pool reflects this admixture of maternal and paternal influences in the development of the Bo’an, Salar, and Dongxiang populations.

Table 5.1 Homozygosity comparison between sub-populations from PR China and Pakistan

PR China sub-populations		**Pakistan** sub-populations	
Bo’an	32.5%	40.9%	***Awan***
Salar	31.5%	28.1%	***Khattar***
Dongxiang	26.7%	28.1%	***Rajpoot***
Mean	**30.23%**	**32.36%**	

(Wang *et al.,* 2000)

In table 5.1, the observed homozygosity levels observed in the Bo’an, Salar and Dongxiang are compared with the three Pakistani communities, which are known to strongly favour consanguineous marriage. For the Pakistani group study, the same ten neutral autosomal markers were used, and it can be seen that these groups possess broadly similar results with respect to homozygosity. The total percentage of alleles shared by the Bo'an, Salar, and Dongxiang communities was 33.1% and the mean percentage of unique alleles per group was 9.61%; this compares with 28.3% alleles shared by the Pakistani communities and 7.7% unique alleles (Wang *et al.*, 2000). These findings indicate that community endogamy, perhaps in combination with preferential consanguineous marriage, is a strongly followed and a long standing practice among Muslim populations in Central, East and South Asia. This would reconcile with reports indicating that in many parts of Asia 20% to 50% of marriages continue to be contracted between closely related individuals (Bittles, 1998).

Of the ten loci investigated, only two showed significant differences in allele size distribution across all three populations ($p < 0.05$), and for the most part mean allele numbers were similar throughout the three populations. These findings add weight to the concept of shared genetic identity at autosomal loci and endogamy within the three communities. The frequency distribution of alleles generally indicated a common pattern among all three communities, with several high frequency shared alleles distributed across all loci (D13S126, 104bp; D13S133, 132bp; D13S192, 103bp; D13s270, 81bp; D15S108, 145bp; D15S11, 243, D15S97, 183bp; D15S98, 157bp; GABRB3, 185bp). These high frequency shared alleles could be old alleles in terms of human evolution, i.e. they may have existed prior to the subdivision of modern human populations. Alternatively, they could have arisen in their most recent common, ancestral generations. The latter explanation might well be the case given the claimed Han background of the founding females in each community.

5.3 Y-chromosome diversity

Phylogenetic analysis was undertaken by tracing gene flow in the Bo'an, Salar and Dongxiang male lineages. As previously described, the construction of a gene tree indicated a greater proximity between the Salar and that of Central Asian, European and NE Han groups, reflecting their Turkic-speaking origins. In contrast, the Bo'an and Dongxiang are both Mongolian-speaking communities and were placed closer to the Mongolian populations. The Hui are an amalgamation of many different populations and also the largest members of the Muslim community; their positioning on the tree reflects a composite gene pool (see figure 4.2).

When the average gene diversities of the study populations were compared with a Hui Muslim community from Liaoning Province, PR China (Black *et al.*, 2000), the Bo'an (0.4230), Salar (0.4610), and Dongxiang (0.4630) communities clearly recorded lower values than the Hui (0.672). The gene diversity of the three study populations may be indicative of political pressures exerted since the establishment of the Peoples Republic Period (1949-present), in particular the policy of autonomous prefectures which resulted in the admixture of certain ethnic minorities. This explanation could account for the high gene diversity exhibited by the Hui, reflecting the governmental decision to group together Muslims who did not possess their own language. Further studies conducted by Black *et al.* (2000) have similarly reported high gene diversity in the majority Han (0.656) population of Liaoning.

The pairwise *Fst* distance was used in this study to assess genetic diversity (table 4.9). The smaller population sizes of the Bo'an and Salar groups may have contributed toward their greater genetic identity. In fact, the numerically larger Dongxiang population have the same Mongolian ancestry as the Bo'an, whilst the Salar are Turkic in origin. As discussed previously, the establishment time afforded to the Dongxiang may have enabled this community to diversify its gene pool. This idea is especially pertinent when considering the founder hypothesis associated with the Y-chromosome lineage. Historical accounts suggest that it was predominantly the transit of males into China, which led to the transfer of culture and social practices and ultimately the formation of ethnic minorities.

Two previously unidentified alleles were observed in the present study (122 bp allele, DYS388 and 99 bp allele, DYS393). At locus DYS388 the allele was present across all three populations, indicating community relatedness that may have evolved as a result of their co-residence and endogamy. Alternatively, the allele at locus DYS393 was only found in the Salar, and may reflect their unique Turkic background. It could also be argued that these alleles are unique genetic variants within each community, and they may have arisen as a result of random mutational events. Further analysis of the variants would, however, need to be carried out to accurately determine their genetic basis.

5.4 Mitochondrial diversity

In terms of effective population size, the Bo'an, Salar and Dongxiang populations are effective genetic isolates in comparison with the Han females and this has allowed mtDNA analysis of maternal lineages which are still fairly cohesive and where extensive admixture is limited. The results of the present study showed characteristics that were consistent with Central Asian populations (Comas *et al.*, 1998). This was evident in two mutations, a C-T transition at position 16223 and a C-T transition at position 16362, identified in the Bo'an, Salar and Dongxiang. Comas *et al.*

(1998) identified a Chinese source of mtDNA in the human migrations along The Silk Road to Central Asia and Europe. This line of reasoning would concur with the idea of males who led a semi-nomadic lifestyle choosing to settle in China and marry Han Chinese females, while others returned to their place of origin possibly to the north or west.

This explanation would support the low mean pairwise nucleotide difference exhibited between individuals in the three groups (5.90, SD = 2.465). The mean pairwise nucleotide difference was examined in relation to other populations, including Middle Eastern, European and Central Asian communities (mean pairwise difference range between 7 and 14) (Comas *et al.*, 1998). The results for the mtDNA sequences indicated a relatively cohesive mitochondrial structure. Due to the isolated nature of the three populations under investigation, it would be anticipated that they would possess a low diversity index, particularly when considered in relation to the highly colonised European and Middle Eastern populations. Another explanation could lie in the long-standing practice of endogamous marriage, which would have the effect of maintaining the suggested Han maternal gene pool over time.

The nucleotide diversity index for the three populations was 0.0162, which was very similar to those reported in the Central Asian Uygurs (0.0164), also a Muslim minority population. Other comparative published nucleotide diversity index results of groups that the male founders are suggested to have originated from include: Eastern Asian (0.0173), Mongolians (0.0180) and Turkish (0.0155) populations. Each of these groups also indicates a relatively cohesive genetic mtDNA sequences. One likely explanation for this lack of variation among the populations may probably have come through various assimilation policies during successive Chinese Dynasties. One such

Imperial decree enforced during the Ming Dynasty (1368-1662 AD) allowed Muslim males only to marry Han Chinese females. Other factors, such as the periods of major civil unrest during the Muslim uprisings of the Qing Dynasty (1662–1908 AD) could have isolated Muslim communities during more recent history. The consequences of assimilation coupled with social and cultural practices may have led to some degree of admixture. However, the civil uprisings would probably have a bottleneck effect on some communities, leading to isolation and an increased founder effect. This would restrict mate choice to the three co-resident communities, i.e. endogamy, which is reflected in the maternal genepool of present day Muslims (Baric *et al.*, 2000a).

5.5 Conclusions

The inter-community genetic differences revealed in the study may reflect the retention by each community of their separate genetic identities through time. Alternatively, they could indicate that, although they live in the same geographical region, unique genetic variants have arisen within each community, varying in frequency across the generations.

Several factors are probably important in maintaining and/or increasing the genetic differences between the three communities. For example, the different languages spoken by the Salar and the Bo'an/Dongxiang would have been a formidable boundary to marriage. In the smaller Salar and Bo'an communities the autosomal data indicated that marriages between close biological kin had been favoured, possibly reflecting some past degree of constraint in marriage partner choice.

The Y-chromosome and mtDNA data were dissimilar, with the Y-haplotypes indicating major differences between the three populations, whereas the mtDNA suggested more cohesive inter-community relationships. This finding is best explained by past inter-marriage of the diverse male founders of each community with Han Chinese females.

5.6 Technical considerations

The Bo'an, Salar and Dongxiang populations were compared with a predominantly European reference populations obtained from GDB, CEPH and Leiden University Forensics Lab. These databases allowed the analysis of shared polymorphisms from a variety of different sample populations. The sharing of variations among human populations is a common feature to all human sub-populations. Inter-community genetic variation (*Fst*) between humans is estimated at 0.1, suggesting that approximately 90% of human genome variation is common to all sub-populations (Sullivan, 1997). Nevertheless, sample information provided by CEPH, GDB and Leiden University Forensics Lab can tend to be ambiguous. For example, the samples conceivably could have been drawn from individuals with a predisposed medical condition, or they may not be randomly selected. Due to the time constraints imposed on this study, multivariate analysis to accommodate these variables into the results was not possible.

These reference population databases have been specifically designed to offer the researcher populations by which to gauge the degree of genetic variation and diversity expected in a random sample population. Researchers in different genetic disciplines use the autosomal and Y-chromosome markers this study has examined because they provide a basis that allows comparisons of their own data with a reference population.

A difficulty experienced using the CEPH and GDB databases with di-, tri- and tetranucleotide markers were the inconsistencies in allele sizes in the published data compared with the study sample results. This difference in allele sizes could be attributed to the use of different primers, e.g. a longer primer sequence would increase the final length of the PCR product identified. Differences also could arise from the use of different fluorescent labelling systems, such as using primers that are labelled during synthesis, as opposed to incorporating the dye into the PCR product during amplification. The latter method being regarded less practical in terms of screening large populations, due to the time involved in preparation, increased handling errors and molecular weight changes.

The allele sizes obtained in this study frequently varied from even to odd values with respect to the published data. For example, at D13S133 the CEPH results are 132bp, whereas the observed value in the study was 131bp. This meant that a subjective decision had to be made as to how best to differentiate between the appropriate allele sizes. The task was made all the more difficult with dinucleotide as opposed to tri- and tetranucleotide results, simply because errors in judgement were more likely with the smaller value repeat alleles than the larger alleles.

Successful marker amplification depended greatly on the correct amplification protocols, thus necessitating regular modification of protocols to obtain the desired product. Optimal primer concentration was consistently identified as being a requisite for successful amplification. In some instances it had been several months since the primers had last been used and some degradation was unavoidable as they naturally breakdown at a rate of approximately 0.1% per month at -20^0C. In most instances, moderately degraded primers posed no major difficulties, once their initial concentrations were increased in the reaction mix. Similarly, the DNA concentrations often needed to be increased, or in extreme cases where the product had totally degraded, fresh extraction of DNA was required. Some DNA products were more prone to template degradation than others, for example the longer DYS19 fragment proved very difficult to amplify in contrast with shorter DYS393 fragment.

Once the PCR product had been successfully amplified, allele detection on the ABI Prism 373 in most instances proceeded smoothly. One obstacle that was encountered involved the increased concentration of PCR product. Following PCR amplification, the exponential nature of this technique occasionally meant that PCR concentrations were too high to permit accurate sizing on the ABI Prism 373. However, further proper dilution of the PCR product overcame this difficulty and allowed the alleles to be successfully identified.

5.7 *Future directions*

The amount of time available to undertake the study was a major limitation that precluded a more extensive analysis of the results. Only ten mtDNA samples from each population were analysed in this study. To perform a thorough examination, the analysis of all 212 samples from the study would have provided a much clearer profile of the maternal gene pool. Furthermore, sample purity placed constraints on the amount of autosomal, Y-chromosome, and mtDNA samples that could be examined. For example, genomic DNA from the blood spots was subject to degradation over time. Continuing research would benefit from fresh whole blood samples, simply because of the overall quality and quantity of available DNA.

Specific attention could also be paid to the historical narrative, as only a limited amount of information was available on the social and cultural aspects of the Bo'an, Salar and Dongxiang, mainly written by a limited number of Western authors. Further research into demographic, anthropological and historical texts published in PR China would be necessary to fully detail the factors that have influenced the formation of these communities. This information should then complement and further clarify the underlying population dynamics of the study populations.

Further analysis of the results should also pay specific attention to the phylogenetic relationships of the populations. In particular, the construction of genetic trees based on autosomal, Y-chromosome and mtDNA loci, thus allowing results to be compared between each population based on maternal and paternal lineages. On a larger scale, reference populations could also usefully be included at all levels to enhance the overall efficacy of the results. Interaction between different distance measures would have the effect of broadening the validity of interpretation of the single measures used. For instance, by using Nei's standard genetic distance in conjunction with delta mu squared and Slatkin's *Rst* distance.

Interpretation of the current research could also be improved by using a complementary set of statistics with, for example, analysis of molecular variance (AMOVA) used to assess genetic variation in the study populations. The hierarchical nature of AMOVA allows a statistically robust assessment of the variance seen in individual chromosomes and populations. The polymorphic information content (PIC) , i.e. the probability that the genotype of a given offspring will identify a marker allele at a locus that was inherited from each parent, is another statistical parameter which could be used to investigate the effect of consanguinity in the study populations. Linkage disequilibrium could also be assessed in the populations, with alleles at different loci that were not in random association statistically assessed to determine any associations.

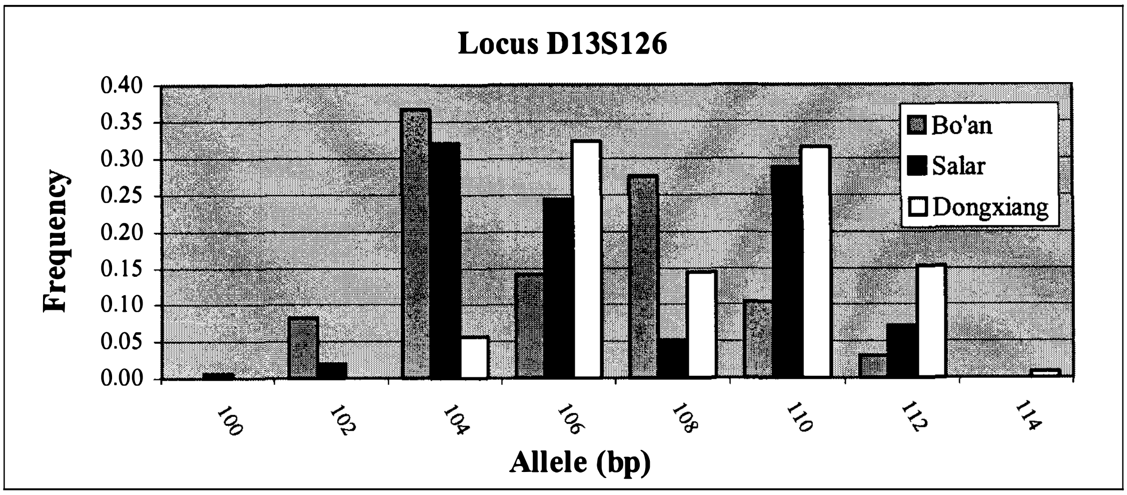
The study of genetic diversity is not just a record of human migrations as in this study, but also details the basis of heritable variation in disease susceptibility. A major future direction for population genetic studies is in the analysis of the most common type of genetic variation, the single nucleotide polymorphism (SNP). There has been growing recognition that large collections of mapped SNPs would provide a powerful tool for human genetic studies. Screening for SNPs in the present study could be used to additionally investigate the diverse genetic histories identified within and between the Bo'an, Salar, and Dongxiang communities. Other applications of SNP analysis include its increased use to improve the resolution of genetic maps. Additionally, SNPs can serve as genetic markers for identifying linkage disequilibrium in isolated populations and can be used to conduct association studies of families with specific disease genes. Although individual SNPs are less informative than currently used genetic markers, they are more abundant and have greater potential for high throughput automation e.g. via Matrix Assisted Laser Desorption/ionisation time-of-flight mass spectrometer (MALDI-TOF MS) (Baric *et al.*, 2000b).

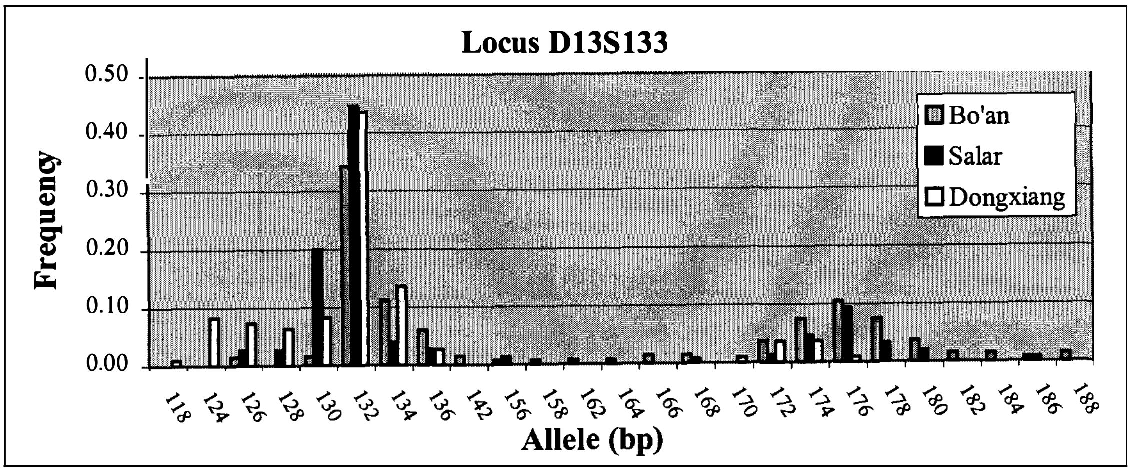
In conclusion, incorporating other Muslim minority populations from PR China, the Central Asian republics and Indo-Pakistani communities would enrich continuing research. Studies conducted by Bittles (1994,1998) and Wang *et al.* (2000) have identified Indo-Pakistani communities as being stratified and highly endogamous. This would allow comparisons and more detailed analysis on the effects of inbreeding in the communities and the possible heritable consequences i.e. segregation of disease genes. To undertake such a detailed analysis of the variation associated with inbreeding or disease genes, family pedigrees would need to be obtained to complement the less informative random samples. Alternatively, increased sample size, could include families from geographically distant regions, and from different ethnic groups, to gain additional insights into whether inbreeding produces comparable homozygosity increases and patterns of autozygosity in all human populations.

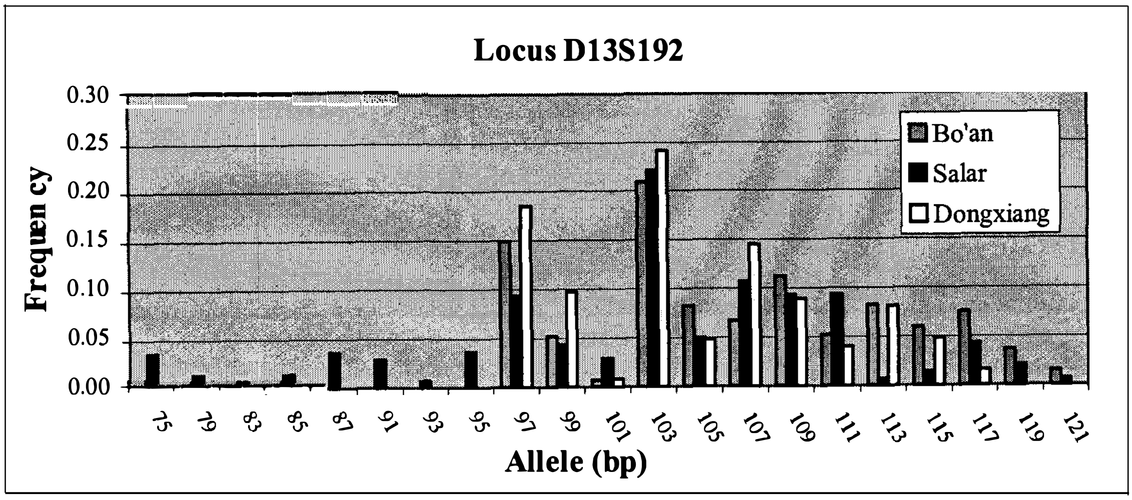
Appendix A - Stanford Microsatellite Markers

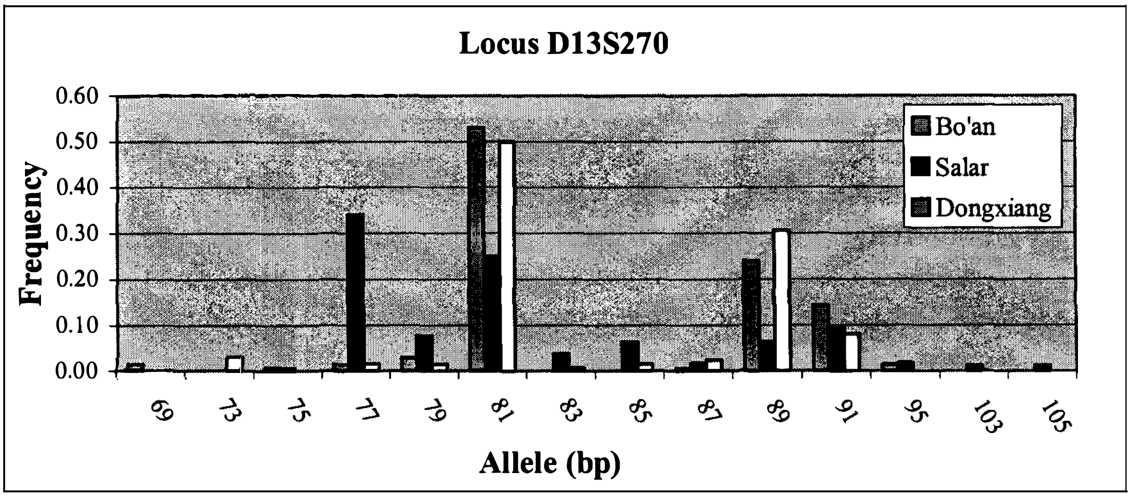
Locus	*Label*	*Oligo Name*	*Sequence*
D13S126	FAM	1303L	TCACCAGTAAAATGCTATTGG
		1303R	GTGATTTTCAAATTTGCTCTG
D13S133	TET	CA008L	GGCAACATAGGGAAACCCTAGC
		CA008R	GCTAGGACTACAGGTGCAAACC
D13S192	HEX	HKCA3-1	GGGTAACATAGCAAGACCCC
		HKCA3-2	AGGTATGAGCCATCTCGTCC
D13S270	HEX	084xc5a (CA)	AGTGCCTGGGTATGAACGTG
		084xc5m (GT)	CTGGAAATGCCTTGGAAGGA
D15S101	HEX	MS178L	GAGCCAAGATCATGTTGC
		MS17R	TGCCCACTAGTTTGAGACA
D15S108	HEX	MFD102L	ATTCTTAACAGGAAGTGAGGG
		MFD102R	AACATGAGTTTCAGAGGGG
D15S11	TET	D15S11L	GACATGAACAGAGGTAAATTGGTGG
		D15S11R	GCTCTCTAAGATCACTGGATAGG
D15S97	FAM	MS14L	TCTCCCTCCAATAATGTGAC
		MS14R	TGAGTCAATGATTGAAATTACTG
D15S98	HEX	MS112L	CATGTGAAACTGCAAAAGCTG
		MS112R	AAAAGTCGCATTTGGTCGTT
GABRB3	HEX	L	CTCTTGTTCCTGTTGCTTTCAATACAC
		R	CACTGTGCTAGTAGTTCAGCTC

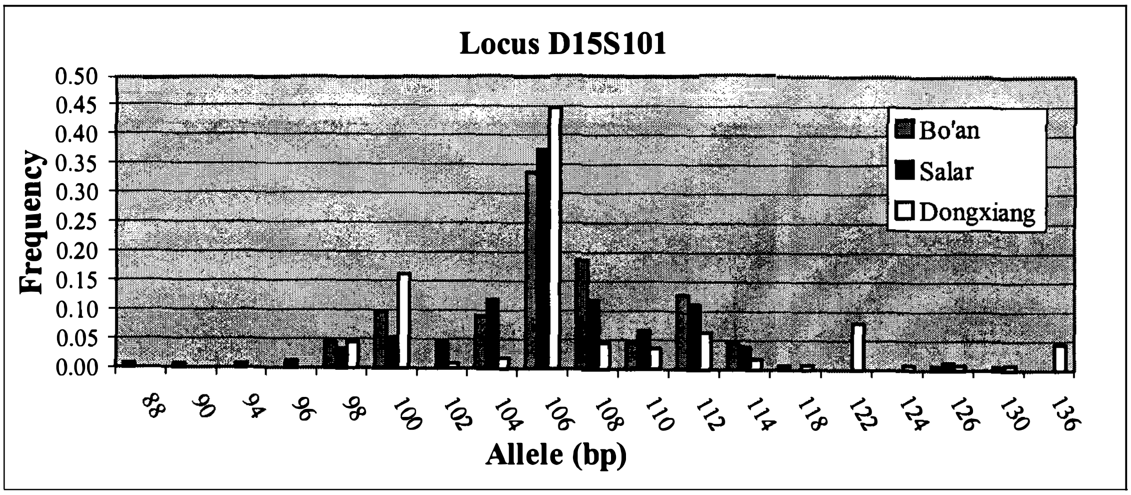
Appendix B - Autosomal allele frequency distributions

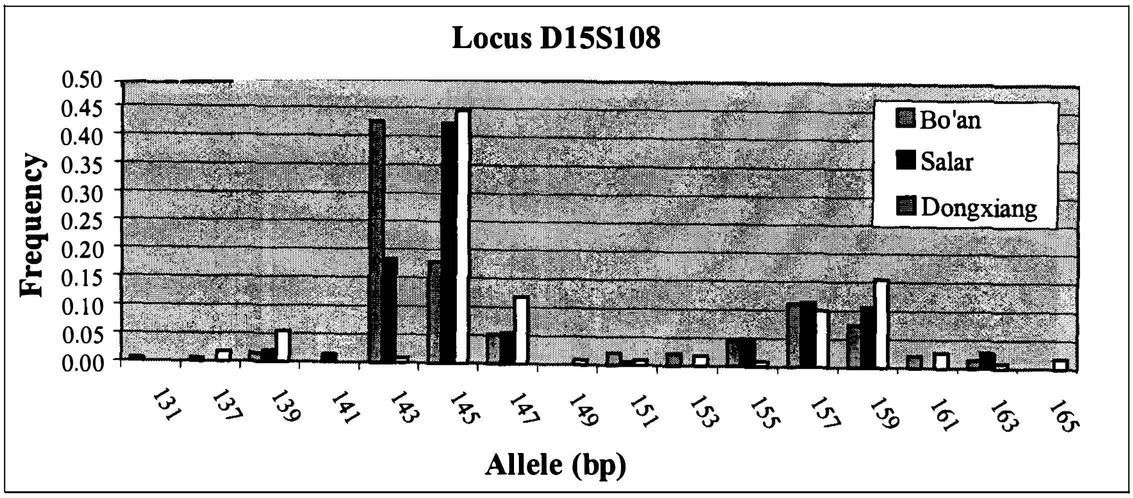


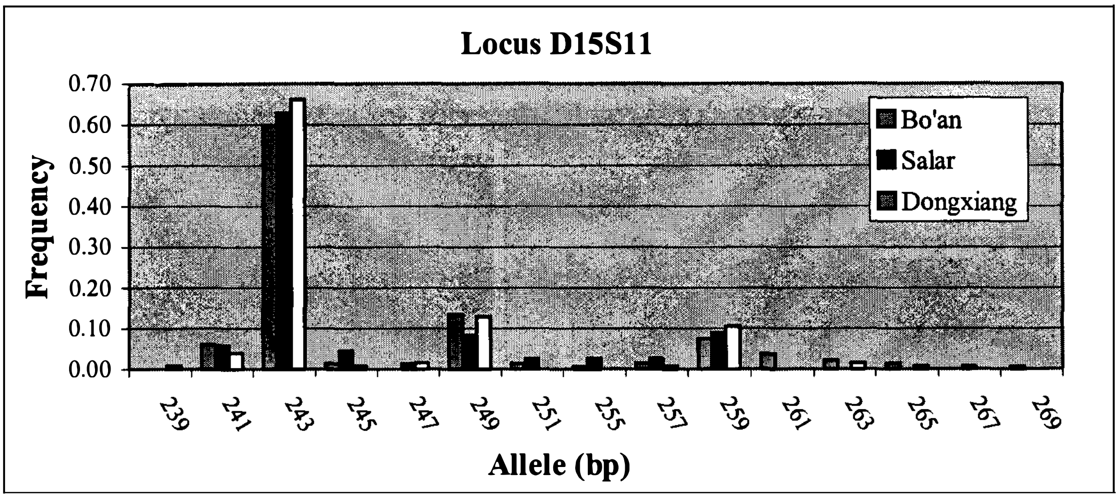


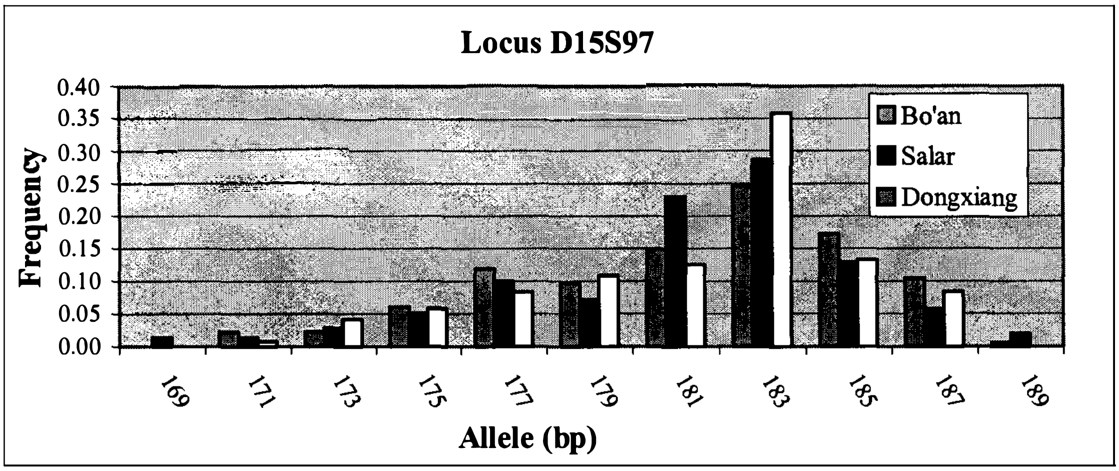


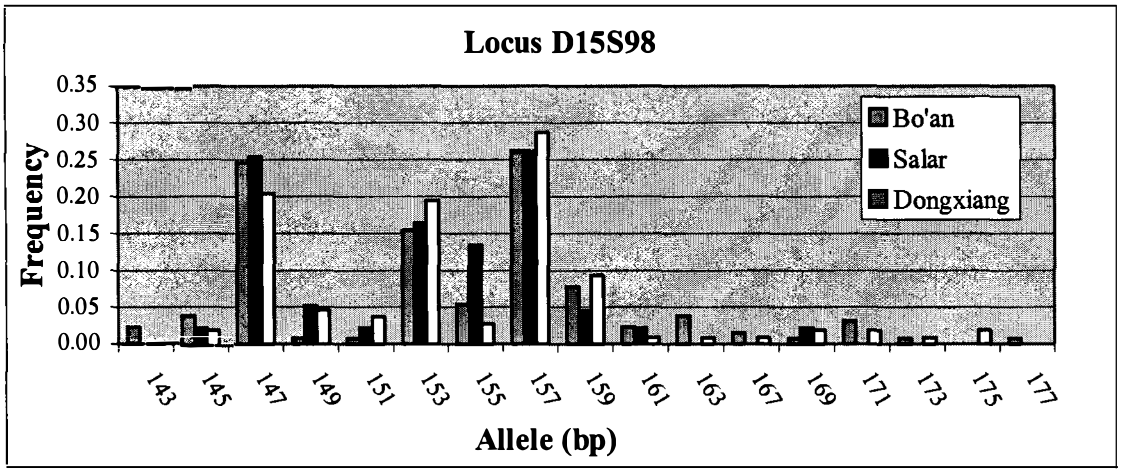

Locus D13S270
Frequency
0.60
0.50
0.40
0.30
0.20
0.10
0.00
69
73
75
77
79
81
83
85
87
89
91
95
103
105
Allele (bp)
Bo'an
Salar
Dongxiang

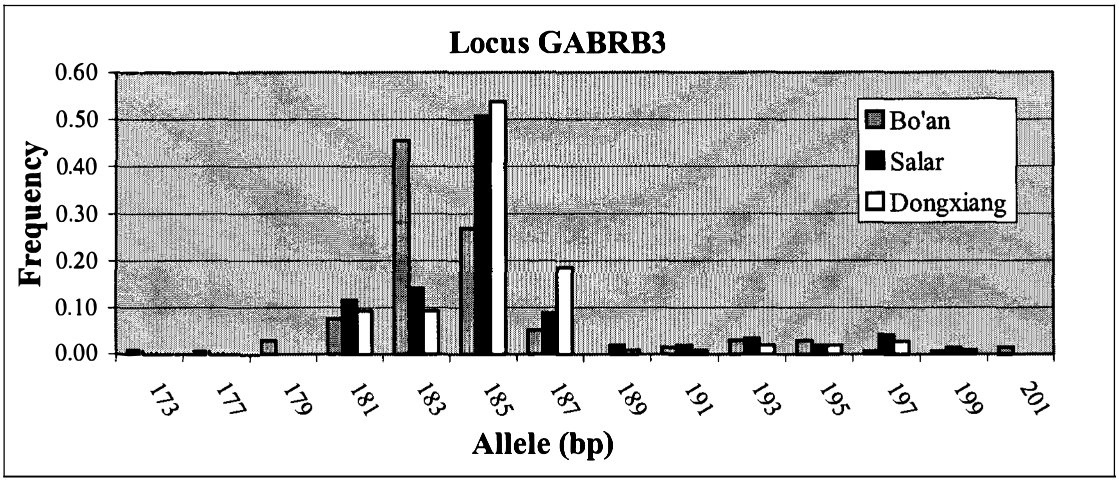

Locus D15S101
Frequency
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
88
90
94
96
98
100
102
104
106
108
110
112
114
118
122
124
126
130
136
Allele (bp)
Bo'an
Salar
Dongxiang


Locus D15S108
Frequency
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
131
137
139
141
143
145
147
149
151
153
155
157
159
161
163
165
Allele (bp)
Bo'an
Salar
Dongxiang


Locus D15S11
Frequency
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Bo'an
Salar
Dongxiang
239
241
243
245
247
249
251
255
257
259
261
263
265
267
269
Allele (bp)


Locus D15S97
Frequency
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Bo'an
Salar
Dongxiang
169
171
173
175
177
179
181
183
185
187
189
Allele (bp)


Locus D15S98
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Bo'an
Salar
Dongxiang
143
145
147
149
151
153
155
157
159
161
163
167
169
171
173
175
177
Allele (bp)


Locus GABRB3
Frequency
0.60
0.50
0.40
0.30
0.20
0.10
0.00
▨ Bo'an
■ Salar
□ Dongxiang
173
177
179
181
183
185
187
189
191
193
195
197
199
201
Allele (bp)

Appendix C

Autosomal expected/observed heterozygosity and _Fis_ values for the study populations

Locus		Bo'an			Salar			Dongxiang	
	Hetero exp	Hetero obs	_Fis_*	Hetero exp	Hetero obs	_Fis_*	Hetero exp	Hetero obs	_Fis_*
D13S126	50.72	47	0.0739	58.62	60	-0.0237	46.83	45	0.0394
D13S133	56.61	42	0.2595	64.34	53	0.1772	42.44	43	-0.0133
D13S192	59.99	56	0.0669	63.90	55	0.1401	53.60	54	-0.0008
D13S270	43.23	43	0.0054	66.03	40	0.3958	40.45	42	-0.0387
D15S101	54.86	47	0.1442	61.74	58	0.0609	42.73	48	-0.1246
D15S108	51.44	37	0.2823	54.13	41	0.2438	41.60	36	0.1356
D15S11	41.41	43	-0.0386	61.93	41	0.3394	33.29	33	0.0089
D15S97	57.08	56	0.0190	63.47	49	0.2293	48.85	46	0.0588
D15S98	54.57	49	0.1027	55.06	46	0.1656	44.90	40	0.1100
GABRB3	47.86	32	0.3330	52.01	61	-0.1742	35.89	39	-0.0876
Total Mean	_51.78_	_45.2_	_0.1248_	_60.12_	_50.4_	_0.1554_	_43.06_	_42.6_	_0.0088_

* Weir and Cockerman (1984)

Appendix D

Heterozygote deficit

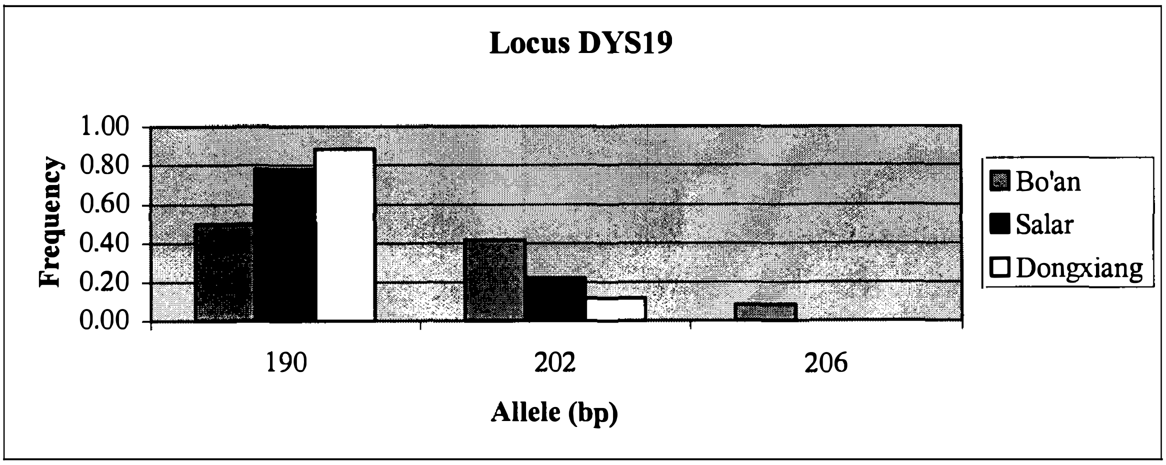
Locus	Bo'an		Salar		Dongxiang	
	p-value	S.E.	*p*-value	S.E.	*p*-value	S.E.
D13S126	0.1067	0.0075	0.0804	0.0090	0.5343	0.0175
D13S133	0.0000	0.0000	0.0000	0.0000	0.1849	0.0186
D13S192	0.0817	0.0122	0.0000	0.0000	0.7687	0.0192
D13S270	0.0175	0.0065	0.0000	0.0000	0.0000	0.0000
D15S101	0.0083	0.0027	0.3771	0.0271	0.1019	0.0210
D15S108	0.0022	0.0013	0.0016	0.0010	0.0000	0.0000
D15S11	0.0268	0.0088	0.0003	0.0003	0.4398	0.0314
D15S97	0.4492	0.0203	0.0553	0.0107	0.0935	0.0111
D15S98	0.0000	0.0000	0.0009	0.0005	0.1807	0.0284
GABRB3	0.0014	0.0008	0.9682	0.0075	0.7473	0.0254
Mean	*0.0694*		*0.1484*		*0.3051*	

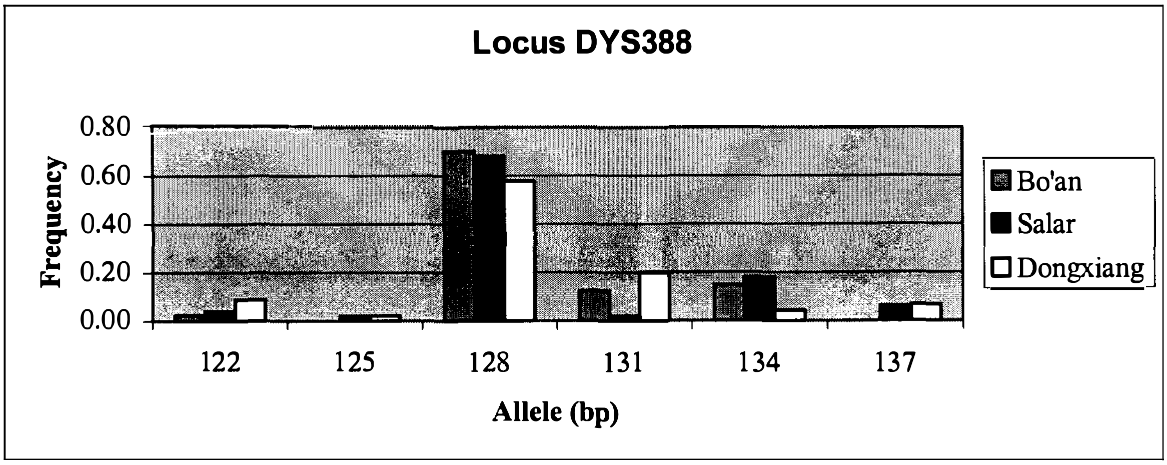
Appendix E

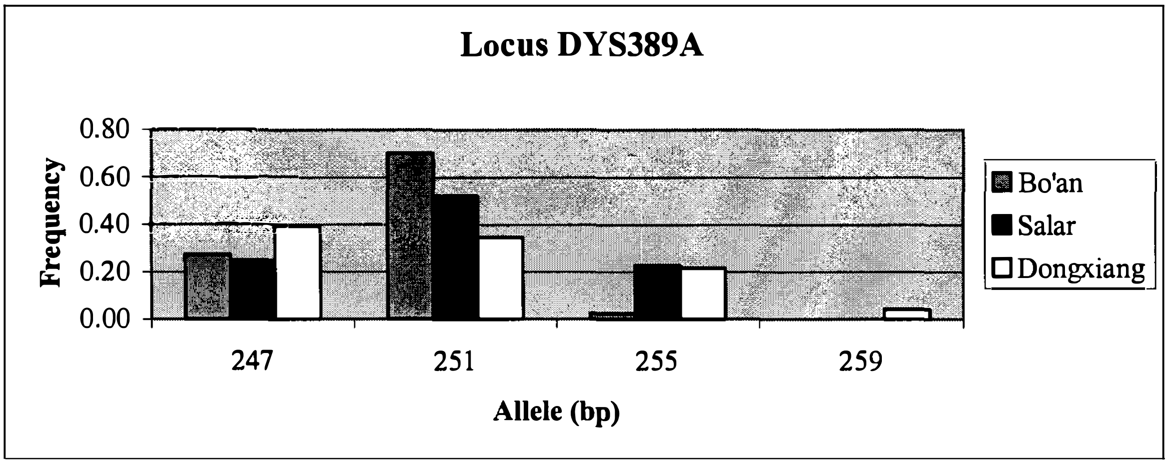
Heterozygote excess

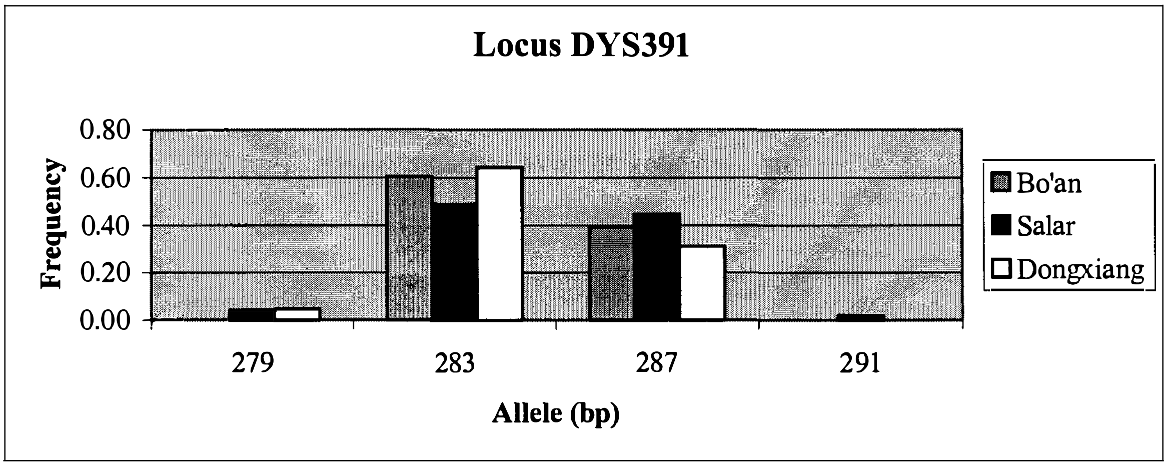
Locus	Bo'an		Salar		Dongxiang	
	p-value	S.E.	*p*-value	S.E.	*p*-value	S.E.
D13S126	0.8846	0.0071	0.8866	0.0099	0.4322	0.0171
D13S133	1.0000	0.0000	1.0000	0.0000	0.7763	0.0255
D13S192	0.8936	0.0162	1.0000	0.0000	0.2590	0.0187
D13S270	0.9911	0.0027	1.0000	0.0000	1.0000	0.0000
D15S101	0.9673	0.0097	0.5990	0.0318	0.8698	0.0239
D15S108	0.9993	0.0007	0.9997	0.0003	1.0000	0.0000
D15S11	0.9965	0.0020	0.9958	0.0022	0.6089	0.0327
D15S97	0.5169	0.0234	0.9592	0.0068	0.9170	0.0100
D15S98	0.9943	0.0033	0.9991	0.0007	0.8720	0.0227
GABRB3	1.0000	0.0000	0.0286	0.0069	0.2402	0.0266
Mean	*0.92436*		*0.8468*		*0.69754*	

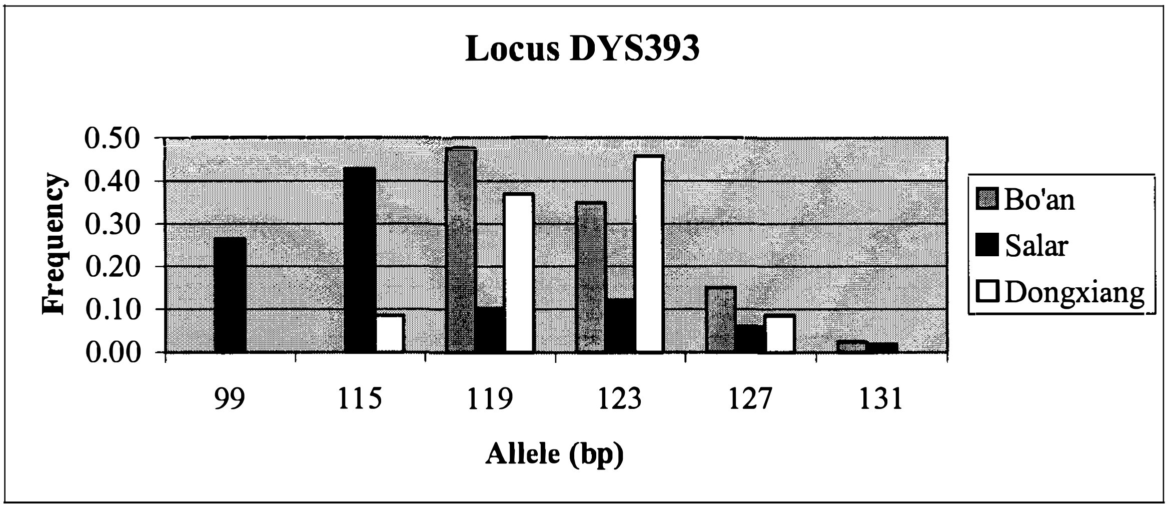
Appendix F - Y-chromosome allele frequency distributions





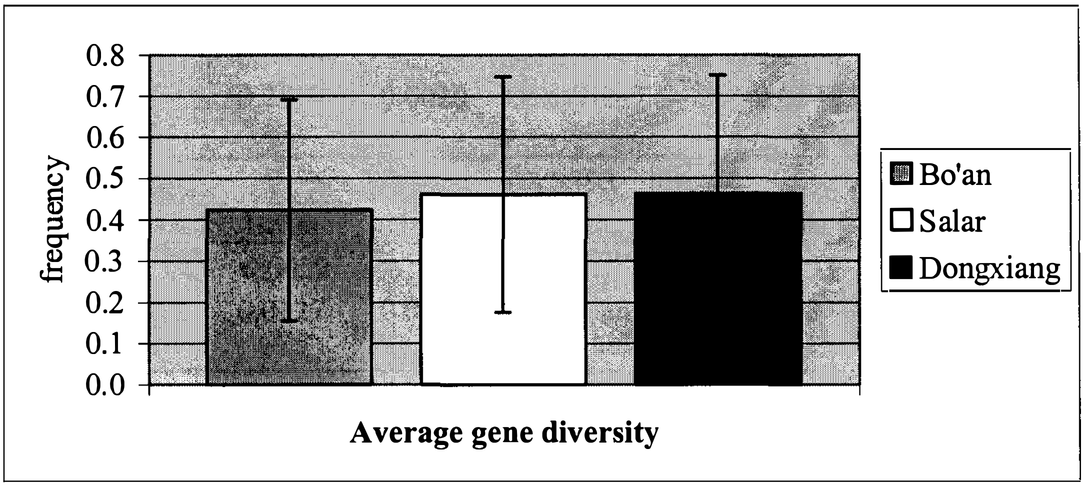



Locus DYS391
Frequency
0.80
0.60
0.40
0.20
0.00
279
283
287
291
Allele (bp)
▨ Bo'an
■ Salar
□ Dongxiang

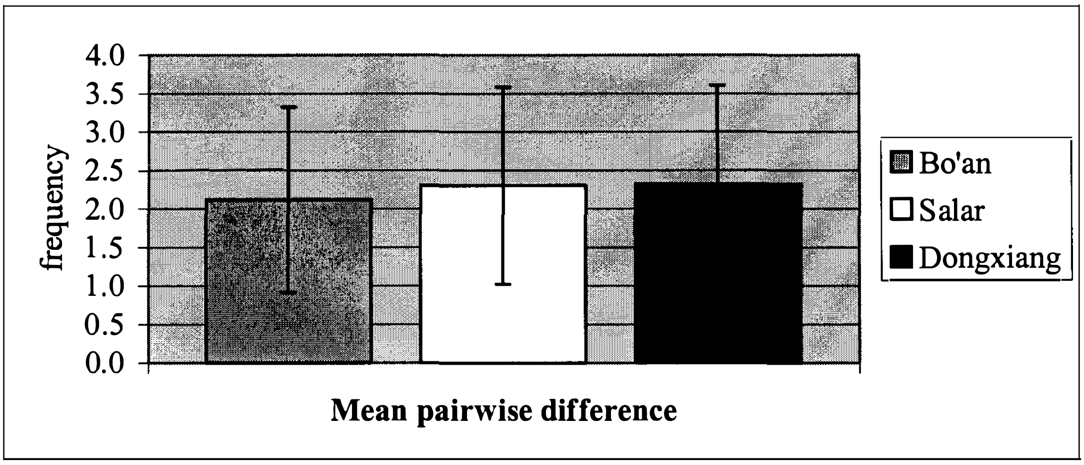

Locus DYS393
Frequency
0.50
0.40
0.30
0.20
0.10
0.00
99
115
119
123
127
131
Allele (bp)
▨ Bo'an
■ Salar
□ Dongxiang

Appendix G

Y-chromosome average gene diversity



Mean pairwise diversity index



Appendix H

Y-chromosome mean pairwise *Fst* distances for all loci

DYS19

	Bo'an	Salar
Salar	0.0830	
Dongxiang	0.2130	0.0020

DYS391

	Bo'an	Salar
Salar	-0.0066	
Dongxiang	-0.0142	0.0170

DYS388

	Bo'an	Salar
Salar	-0.0059	
Dongxiang	0.0127	0.0341

DYS393

	Bo'an	Salar
Salar	0.2333	
Dongxiang	0.0033	0.1977

DYS389A

	Bo'an	Salar
Salar	0.0418	
Dongxiang	0.1147	0.0173

All loci mean

	Bo'an	Salar
Salar	0.0691	
Dongxiang	0.0659	0.0536

Appendix I – Y-chromosome microsatellite markers

Locus	*Fragment size*	*Sequence*
DYS19	186bp	(F) 5'-ctactgagtttctgttatagt-3' (R) 5'-atggcatgtagtgaggaca-3'
DYS388	143bp	(F) 5'-gtgagttagccgtttagcga-3' (R) 5'-cagatcgcaaccactgcg-3'
DYS389A	255bp	(F) 5'- ccaactctcatctgtattatctatg-3' (R) 5'-tcttatctccacccaccaga-3'
DYS391	279bp	(F) 5'-ctattcattcaatcatacaccca-3' (R) 5'-gattctttgtggtgggtctg-3'
DYS393	131bp	(F) 5'-gtggtcttctacttgtgtcaatac-3' (R) 5'-aactcaagtccaaaaaatgagg-3'

References

Anderson, S., Bankier, A.T., Barrell, B.J., deBruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R, & Young, I.G. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-465.

Baric, T., Wang, W. and Bittles A.H. (2000a). Genetic diversity among the Salar, Bo'an, and Dongxiang Muslim populations in the Peoples Republic of China. *Journal of Medical Genetics* **37**, Supp 1.

Baric, T., Worsley, P., Wang, W., Kalaydjieva, L., & Bittles, A.H. (2000b). MALDI-TOF mass spectrometry and single nucleotide polymorphisms in human populations. *Journal of Medical Genetics* **37**, Supp 1.

Bittles, A.H. (1994). The costs of human inbreeding and their implications for variation at the DNA level. *Nature Genetics* **8**, 117-121.

Bittles, A.H. (1998). *Consanguineous marriage: empirical estimates of the current Global prevalence and their outcomes. Working paper number 74.* Stanford: Morrison Institute for Population and Resource Studies, Stanford University.

Black, M.L., Wang, W., and Bittles, A.H. (2000). A genome-based study of the Muslim Hui community and the Han population of Liaoning Province, PR China. *The Annals of Human Biology* (in submission).

Bowcock, A.M., Ruiz Linares, A., Tomfohrde, J., Mich, E., Kidd, J.R., & Cavalli-Sforza (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**, 455-457.

Charlesworth, B., Sniegowski, P., & Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**, 215-220.

Coltman, D.W., Bowen, W.D., and Wright, J.M. (1998). Birth weight and neonatal survival of harbour seal pups are positively correlated with genetic variation measured by microsatellites. *Proceedings of the Royal Society London, Series B* **265**, 803-809.

Comas, D, Calafel, F., Mateu, E., Pèrèz-Lezaun, A., Bosch, E., Martinez-Arias, R., Clarimon, J., Facchini, F., Fiori, G., Luiselli, D., Pettener, D., & Bertranpetit, J. (1998). Trading genes along The Silk Road: mtDNA sequences and the origin of Central Asian Populations. *American Journal of Human Genetics* **63**, 1824-1838.

De Knijff, P., Kayser, M., Caglia, A., Corach, D., Fretwell, M., Gehrig, C., Graziosi, G., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterriech, W., Pandya, A., Parson, W., Penacino, G., Pèrèz-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P.M., Szibor, R., Teifel-Greding, J., Weichhold, G., & Roewer, L. (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. *International Journal of Legal Medicine* **110**: 134-149.

Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M., & Friemer, N.B. (1994). Mutational processes of simple sequence repeat loci in human populations. *Proceedings of the National Academy of Science USA*. **91**, 3166-3170.

Du, R. & Yip, V.F. (1993). *Ethnic groups in China.* Science Press: Beijing and New York.

Ebrey, P.B. (1999). Cambridge illustrated history China. Cambridge University Press: London.

Esteban, P., Nilmani, S., Agustinus, G. Soemantri, S.T., McGarvey, J., Hundreiser, M.D., Shriver, R., & Ranjun, D (1999). Genetic variation at 9 autosomal microsatellite loci in Asian and Pacific populations. *Human Biology* **71**, 757-768.

Family Planning Commission (1998). *Chinese Family Planning Yearbook 1997.* Family Planning Commission: Beijing.

Gladney, D.C. (1998). *Ethnic identity In China: The making of a Muslim minority nationality.* Harcourt Brace & Company: Fort Worth, Texas.

Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L., Feldman, M.W. (1995). An evaluation of genetic distances for using microsatellite loci. *Genetics* **139**, 463-471.

Hartl, L.D., & Clarke, A.G. (1997). *The principles of population genetics* (3rd Ed). Sinauer Associates, Inc: Massachusetts.

Jarne, P. & Lagoda, P.J.L. (1996). Microsatellites, from molecules to populations and back. *Tree* **11** , 424-429.

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

Kimura, M., & Weiss, G.H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561-574.

Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I., & Khorana, H.G. (1971). Studies on polynucleotides XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of Molecular Biology* **56**, 341-361.

Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). *Molecular Cloning, A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

McBride, L.J., Koepf, S.M. Gibbs, R.A., Salser, W., Mayrand, P.E., Hunkapiller, M.W. and Kronick, M.N. (1989). Automated DNA sequencing methods involving polymerase chain reaction. *Clinical Chemistry* **35**, 2196-2201.

Murray, B.W. (1996). The estimation of genetic distance and population substructure from microsatellite allele frequency data. At: http://helix.biology.mcmaster.ca/brent/node1.html.

Nei, M. (1972). Genetic distances between populations. *American Naturalist* **106**, 283-292.

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

Resnik, D. B. (1999). The Human Genome Diversity Project: ethical problems and solutions. ***Politics and the Life Sciences*** **18**, 15-24.

Rahman, Y. A. (1997). Islam in China.
At: http://Salarm.muslimsonline.com/~azahoor/islchina.htm

Peréz-Lezuan, A., Calafel, M., Seielstad, E., Mateu, E., Comas, D., Bosch, E., & Bertranpetit, J. (1997). Population genetics of Y-chromosome short tandem repeats in humans. *Journal of Molecular Evolution* **45**, 265-270.

Population reference bureau (2000). *Population data sheets*. Washington D.C.

Roberts, J.A.G. (1999). *A concise history of China*. Harvard University Press: Massachusetts.

Sanger, F., Morel, C., & Cedergren, R.J. (1977). DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* **74**, 5463-5467.

Shriver, M.D., Jin, L., Chakraborty, R., & Boerwinkle, E. (1993). VNTR allele frequency distribution under the stepwise mutation model: A computer simulation approach. *Genetics* **134**, 983-993.

Strachan, T. & Read, A.P. (1999). *Human Molecular Genetics* (2nd ed). Bios Scientific Publishers: Oxford.

Sullivan, S. (1997). *A study of the effects of consanguinity at the genomic level in two Pakistani bradaris*. Edith Cowan University: Perth.

Valdes A.M., Slatkin M., Freimer N.B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737-49.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O., & Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science* **280**, 1540-1542.

Wang, D.G. *et al.* (1998). Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms the human genome. *Science* **280**, 1077-1082.

Wang, W., Sullivan, S. G., Ahmed, S., Chandler, D., Zhivotovsky, L.A., & Bittles, A.H. (2000). *Annals of Human Genetics* **64**, 41-49.

Weber J.L., & Wong C. (1993). Mutation of human short tandem repeats. *Human MolecularGenetics* **2**, 1123-1128.

Weir, B.S. and Cockerham, C.C. (1984). Estimating F-statistic for the analysis of population structure. *Evolution* **38**, 1358-1370.

Wong, H.M. & Dajani, A.A. (1988). *Islamic frontiers in China.* Scorpion: London.

Woolf, T.M. & Taylor, M. (1998). Studies of populations and genetic diseases: mixing it up. *Trends In Genetics* **14** , 218-219.

Wright, S. (1921). Systems of mating. *Genetics* **6**, 111-178.