1-1-2000

Analysis of the population genetics of the Han and Hui of Liaoning province, Peoples Republic of China

Michael L. Black
Edith Cowan University

Edith Cowan University

Copyright Warning

USE OF THESIS

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

Analysis of the population genetics of the Han and Hui of Liaoning Province, Peoples Republic of China.

by

Michael L Black

Student No 3890797

A Thesis Submitted in Partial Fulfilment of the Requirements for the Award of Master of Science.

At the Faculty of Communication and Health Sciences, Edith Cowan University, Joondalup.

Date of submission: 20/12/1999

ABSTRACT

Throughout recorded Chinese history, regions of the country populated by persons of non-Han ancestry often fluctuated significantly in population numbers and in their political and commercial influence. However, at all times they were considered as important contributors to the nation. Many of these peoples had moved from their homelands, settled in China and had intermarried with Han Chinese. Over the generations they became accepted as fully-fledged Chinese citizens although, in many instances, they retained their traditional customs and religious practices, and frequently their own language. The Hui Muslims are a good example of this process of integration, and today they comprise some 8.6 million individuals thus forming approximately half of the total Muslim population of PR China.

The purpose of this study is to investigate the genetic structure of two populations, the Han and Hui of Liaoning, Northeast PR China. The study seeks to provide a better understanding of the effect of population subdivision on the genetic diversity of human populations, by comparing genome-based investigations using single tandem repeat markers with historical and anthropological information. As the Hui of Liaoning are endogamous, and they are known to contract consanguineous marriages, the study also attempts to assess the effect of consanguinity on overall genetic diversity in the Hui.

Genetic analysis of the Han and Hui was undertaken by surveying the allele distribution patterns at ten autosomal and seven Y-chromosome microsatellite loci in both study populations. Various population genetic

techniques were applied, based either on the Infinite Allele Mutation model or the Stepwise Mutation Model. It was found that both the Han and the Hui exhibited appreciable heterogeneity at autosomal and Y-chromosome loci, indicative of the presence of population substructure and that the AMOVA test best defined genetic relationship between two populations. It was concluded that further detailed anthropological and demographic information was needed to provide a more detailed account of population structure and for the creation of a detailed phylogeny tracing male Hui gene flow.

It was also found that consanguinity seemed to have a negligible effect on the genetic diversity of the Hui population of Liaoning. It was concluded that either the practice of consanguinity had not occurred over a sufficiently long time period to alter overall genetic diversity or that heterozygote advantage may be operating at various loci.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher education; and that to the best of my knowledge and belief it does not contain any material previously written by another person except where due reference is made in the text.

Signature_

Date 31/3/2000

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisors. I thank Dr Wei Wang for his guidance and encouragement and Professor Alan Bittles for his boundless enthusiasm and faith in my ability to complete this work.

I would also like to thank David Chandler, Salvatore Di Grandi and everyone else at the Centre for Human Genetics for their support and friendship.

I thank those involved in the collection of samples used in this study especially H.L. Jia from the Peoples Liberation Army No. 201 Hospital, Liaoyang and C. Qian of the China Medical University, Shenyang and the Han and Hui of Liaoning province, P.R. China, without whose involvement this study would not have been possible.

Finally, I would like to thank my family for their support and encouragement.

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Human populations and the concept of race

The concept of race was first applied in the study of human populations in the eighteenth century with the aim of extending to humans a taxonomic classification below the level of species (Senior and Raj 1994). As time progressed, the study of human variation reflected general sociopolitical biases derived from human social experience that carried over to the scientific realms (Lewontin 1972).

The publication of Darwin's *Origin of the Species* gave rise to a belief that Northern European man was more fully evolved than any of the other races (Marks 1995). Thus began numerous studies based on phenotypical traits such as skull size, skin colour, height, eye shape and intelligence. These phenotypic studies bloomed with the growth of eugenics movements, which reached their peak in the 1920s and 1930s (Marks 1995).

It can be argued that the first studies into human genotypic, as opposed to phenotypic, variation began in the first decades of the 20th century, when variation in the ABO blood grouping system amongst pre-conceived human races was demonstrated (Weiss 1998). The study of the racial variation of ABO blood groups was conducted by Hirszfeld and Hirszfeld (1918-19) with four ABO "types" identified, European, Asian, African and Intermediate, based on the ratio of the blood groups A and AB to blood groups B and AB. Later studies showed this was simply the result of the eugenic bias of the researchers, as distinct racial boundaries in the distribution of ABO genotypes did not exist (Marks 1995). Instead, what was apparent was that human blood group data showed a gradual change in the frequency of specific genotypes across space (clinal variation),

rather than exhibiting clear boundary differences between groups (Cavalli-Sforza *et al.* 1994).

1.2 Protein polymorphism

It was not until the 1960s that the extent of genetic variation in the human species was appreciated, concomitant with the introduction of protein electrophoresis. It was found that a gene could have many functionally equivalent alleles, seemingly in direct contradiction to the classical Mendelian model of mutational deviations evolving from one wild type allele. This situation has been termed neutral variation and was first defined by Kimura (1968).

In the 1970s many studies concentrating on worldwide human diversity based on blood groups and protein families were published, for example Lewontin (1972) and Cavalli-Sforza and Bodmer (1971). These studies centred on genotypes based on blood groups, allozyme data, and other protein families. The results used to estimate times since the separation of major racial (continental) populations, by treating groups at the extremities of the human species distribution as if they had been living in complete isolation and estimating how long the present differences between them would have taken to accumulate (Weiss 1998). All studies found most variation within African populations, lending support to the Out of Africa theory of human origins, initially developed from the fossil record.

Lewontin (1972) came to the conclusion that only 6.3% of human diversity was assignable to race. That is, most variation, (93.7%) is due to differences between individuals within a particular racial group. He therefore concluded that racial classifications were of no genetic or taxonomic significance and so no justification could be offered for their continued usage.

1.3 DNA polymorphism

Studies on variation using protein polymorphisms are, by definition, limited, as proteins only represent that fraction of the human genome that encodes functional products. It is now estimated that approximately 10% of the human genome contains functional genes, with the rest often described as junk DNA, mainly consisting of sequences such as pseudogenes, long interspersed sequences (LINES), short interspersed sequences (SINES) and satellite DNA.

The first method developed for quantification of the molecular variation of DNA was Restriction Fragment Length Polymorphisms or RFLPs. RFLPs are fragments of DNA produced through cleavage of DNA strands by restriction enzymes. Restriction enzymes cleave DNA at specific nucleotide sequence recognition sites e.g., CAAG, resulting in the production of DNA sequences that vary in size according to the placement of the restriction site. Variation occurs between individuals due to mutations at restriction sites either disabling the occurrence of restriction sites or creating new sites, in turn creating new DNA sequence variants. These restriction enzyme cleavage sites are detected by Southern blot hybridisation. RFLP analysis was soon utilised in human population genetic analysis. An example is the study by Wainscoat (1984) where RFLPs in the β-globulin gene cluster were analysed in eight different human populations. The result was a cladistic lineage similar to those produced from earlier protein polymorphism studies.

The use of RFLP technology in human mitochondrial sequences by Cann, Stoneking and Wilson (1987) resulted in the mitochondrial Eve hypothesis which calculated that the ancestral human female could be dated back 200,000 years. The same study also showed that it was difficult to locate any distinctive

geographical patterns amongst the human samples. Individuals from the same local population generally had similar mtDNAs, but there often was evidence that persons from other different populations were scattered amongst them.

The discovery of variable number of tandem repeat loci (VNTRS), or minisatellite DNA, by Jeffreys *et al.* (1985) led to the creation of a tool with which geneticists could resolve identification of the individual for forensic purposes. This concept, termed DNA fingerprinting, further demonstrated that the great majority of human genetic variation could be ascribed to differences between individuals and not to differences between groups.

The subsequent discovery of smaller repeat sequences termed microsatellites (Weber and May 1989) resulted in the development of what is currently the most widely used genetic marker. A worldwide survey of indigenous populations, akin to that attempted with RFLP analysis, revealed that microsatellite data showed discrete clusters which corresponded with the population of origin (Bowcock *et al.* 1994). The success of the technique was probably due to the fact that microsatellites exhibit much greater variation than any of the classical polymorphisms and nuclear RFLPs, and they also have greater diversity than mtDNA (Bowcock *et al.* 1994). However, the clustering patterns were enhanced by the fact the samples were collected from geographically discrete populations, and a more randomised sample would probably give a more complex picture.

The most recent development in the study of human genetic variation is the detection of single nucleotide polymorphisms (SNPs). A SNP is a position on the DNA ladder at which two alternative bases occur at appreciable frequency (>1%). They are the most common polymorphism in the human genome

occurring at a density of about 1 every 1kb of DNA (Chakravarti 1998). The development of so called DNA Chip technology will allow the simultaneous scanning of hundreds, even thousands of SNPs in an individual. Utilisation of this technology will add to the ever more complex picture of human genetic variation.

1.4 The concept of ethnicity

The discovery of blood group, protein and DNA polymorphisms indicates that human genetic variation is fluid, created by mutation, but vetted over history by biological, demographic and historical processes (Chakravarti 1998). From this conclusion it is clear that no race possesses a discrete package of genetic characteristics (Senior and Raj 1994), and so classification of the human species on racial grounds cannot be regarded as scientifically valid.

New concepts of human population structure are, however, needed. Unlike race, ethnicity is a socially constructed phenomenon, and ethnic boundaries are, by their very nature, imprecise and fluid (Senior and Raj 1994). The definition of ethnic groups is based on linguistic, religious and cultural differences. In other words, ethnicity is primarily an anthropological construct which has only recently been recognised in human genetics.

1.5 Genetic anthropology

One result of the recognition of ethnicity in the field of human genetics is the development of the new scientific discipline, genetic anthropology. Genetic anthropology can be considered as a merger of population genetics and anthropology. The discipline utilises patterns of genetic similarity among different human populations to infer demographic history, including mating structure, the history of migration and admixture with surrounding groups, and

population size fluctuations, with cultural information such as linguistic characteristics, archaeological artefacts and historical records.

Ambitious projects are now in train where such anthropological data are being collected in combination with genetic information. An example is Eurasia 98 which was planned as a collaborative American-British-Uzbek anthropological expedition to the regions of Transcaucasia, Central Asia and Siberia, to investigate the genetic anthropology of Central Asian populations (see appendix A)

A more ambitious project, called the Human Diversity Database, is being conducted at the Human Population Genetics Laboratory at Stanford University (see appendix A). The Database is planned as a comprehensive community repository supporting work in human population genetics and quantitative anthropology. Initially it will contain data published in the book *History and Geography of Human Genes* (Cavalli-Sforza *et al.* 1989), but it is hoped that a collection of published and unpublished DNA data by individual and by population (including RFLPs, microsatellites, and SNPs), and data from the Centre d'Études du polymorphisme Humaine (CEPH) database will be included. Eventually it is envisaged that the Database will serve as the core source of information derived from the Human Genetic Diversity Project.

1.6 Conclusion

The establishment of large, global collaborative projects highlight the need for a multi-disciplinary approach to understanding human genetic diversity. It is now apparent that the study of human genetic diversity is more historical and medical in orientation than biological and taxonomic (Senior and Raj 1994). Using genetic approaches it is possible to trace population migrations, analyse the effect of population bottlenecks resulting from major historical events (e.g., the

effect of major disease epidemics) and assess the effects of population isolation and consanguinity.

1.7 Aims of the study

The aim of the present study is to adopt a genetic anthropological approach to investigate the structure of two co-resident populations, the Han and Hui of Liaoning, Northeast PR China via the use of historical, demographic and genetic data. In summary, the study will attempt to:

1) Determine if distinct patterns of genetic diversity can be defined between the Han and the Hui populations of Liaoning province at autosomal and Y-chromosome single tandem repeat loci.
2) Determine if, within each population, the patterns of diversity are comparable at autosomal and Y-chromosome loci.
3) Determine if patterns of Y-chromosome allele variation observed in the Hui can be ascribed to male-directed gene flow; and
4) Determine the effect of inbreeding on the observed pattern of genetic variation in the Hui.

Thus it is envisaged that the study will lead to a better understanding of the effect of ethnic affiliation and consanguinity on the genetic diversity of human populations, by comparing genome-based investigations using single tandem repeat markers with historical and anthropological information.

The study will also attempt to compare Y-chromosome and autosomal data, a topic that to date has received lesser attention because of the paucity of appropriate Y-chromosome markers (Roewer *et al.* 1996, Péréz-Lezuan *et al.* 1997b, de Knijff *et al.* 1997). Comparisons between autosomal and Y-

chromosome data should prove useful, as historical references indicate that the main founders of the Hui population were mainly male Arab and/or Iranian traders who married Han women (Wong and Dajani 1988; Du and Yip 1993; Gladney 1998). Thus Hui males should exhibit significantly different Y-chromosome haplotypes than their Han counterparts but share autosomal haplotypes.

Chapter 2

Historical background

2.1 Introduction

This study is based on two ethnic groups, the Han and Hui, who are co-resident in the northeastern province of Liaoning, PR China. In order to adequately analyse genetic variation within and between the two communities, knowledge of the history of the area in which they live and a broad understanding of their individual histories is needed. This information will allow a perspective on the genetic structure of the communities in terms of the effects of population migration, admixture and isolation.

2.2 Chinese Dynasties

In the following discussion of the historical background of this study, much of the historical chronology will be discussed in terms of Dynastic periods as much of Chinese history prior to the 20th century was classified in this way. Thus Table 2.1 shows in chronological order the major Chinese dynasties/periods of the modern era.

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
Song Dynasty	960 AD – 1279 AD
Yuan Dynasty	1279 AD – 1368 AD
Ming Dynasty	1368 AD – 1662 AD
Qing Dynasty	1662 AD – 1908 AD
Republican period	1908 AD – 1949 AD
Peoples Republic period	1949 AD –

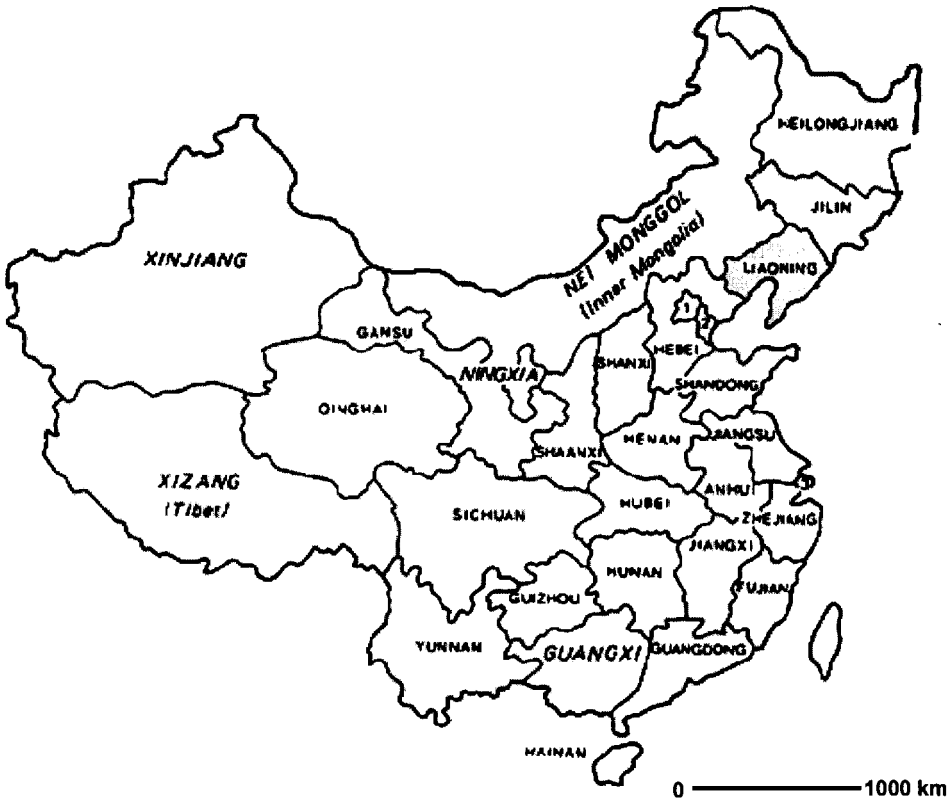
2.3 Liaoning province

Liaoning province is located in Northeast China on the border with North Korea (Figure 2.1), and forms part of the region formerly called Manchuria. Liaoning has a population of approximately 40 million people, of which the Han majority numbers approximately 33 million and the Hui minority 263 000 (Family Planning Commission 1997).

The area that is now the southern tip of Liaoning province was for many centuries the border between the Chinese Empire and the "barbarian" hordes, the Jurchen (Fairbank and Reishauer 1990). Through contact with the Chinese Empire the Jurchen tribes learned of Chinese culture, resulting in the formation of a Manchu nationality and eventually the successful founding of the Qing dynasty in 1662 AD. During the early Qing period Mukden (now modern Shenyang) became the centre of Qing Dynasty power, before the seat of government was moved to Beijing (Fairbank and Reischauer 1990).

Liaoning, as part of the puppet state Manchuko was under Japanese rule from 1932-1945 (Fairbank and Reischauer 1990). It served as a centre for heavy industry with its rich natural resources being used to fuel the Japanese Military machine. Liaoning remains an industrial region, with parts of the province enjoying special economic status under the current government of the Peoples Republic of China.

Figure 2.1 Map of the Peoples Republic of China



**Shaded area is Liaoning province*

2.4 Chinese ethnography

China is usually portrayed as a homogenous monoethnic state, but this is far from the truth (Gladney 1998). In fact, China is a multicultural and ethnically diverse nation with great cultural, geographic, and linguistic diversity among its dispersed populations. The Peoples Republic of China (PR China) is composed of 56 officially recognised nationalities, including the majority Han nationality. The other 55 nationalities have a total population of approximately 91 million (Family

Planning Commission 1997). The largest minority nationality are the Zhuang peoples of Guangxi province who number approximately 15 million, and the smallest are the Luoba of Tibet numbering just 2,312 individuals.

A contemporary definition of ethnicity is that members of an ethnic group share consciousness of group solidarity, by virtue of sharing common descent and common custom and habits (Lipman 1998). The government of PR China gives ethnicity, or in official terminology, *minzu* (nationality), definition by markers such as common territory, language and economy (Gladney 1998).

Chinese history sheds light on this definition of ethnicity. The concept of the "Middle Kingdom" was central to Chinese thinking, and probably still is (Fairbank and Reischauer 1990). Throughout Chinese history, the Chinese ruling elite viewed China as the middle kingdom, the centre of the world with all other states in the "barbarian" fringe. Foreign "barbarians" were labelled *wayji* (outside barbarians), whereas minorities who lived within the middle kingdom were labelled *neiyi* (inside barbarians) (Dikötter 1992). This practice is in accordance with the Confucian practice of *zheng ming* (rectification of names), whereby labelling and categorisation restores order and all is well with the world (Leslie 1986).

Further categorisation of populations within China was also a common practice of the ruling dynasties. One example is the categories defined by the Mongol dynasty (1279-1368), with the population of imperial China divided into four categories: The Mongols, *Semu* ('the coloured eyes', Western and Central Asians), *Hanren* ('Han people', Northern Chinese, Khitans , Jurchens and Koreans), and *Nanren* ('southerners') (Dikötter 1992). In essence, by categorising populations, a sense of order was created which allowed the ruling

dynasties to create the impression of firm control over their dominions. The modern-day official *minzu* is little different in that it seems to create a sense of order out of the complex ethnography of PR China, thus giving the government at least the appearance of central control over the populace.

2.5 The Han *minzu*

The Han can trace their history back to the Huaxia period that extended from the 21st to the 8th centuries BC. In essence, it is believed that the Han were an ethnic group based on the ancient Huaxia of the middle and lower reaches of the Yellow River, who subsequently assimilated other local and regional ethnic groups (Du and Yip 1993).

The term Han is used collectively to define the majority population in PR China. It was adopted into common usage after the fall of the Eastern Han Dynasty which ruled between 25 AD and 220 AD, often thought of as a golden age in Chinese history (Du and Yip 1993). According to the 1990 census, the Han make up 92% of the population of PR China and number approximately 1,100 million (Family Planning Commission 1997). The Han are resident throughout PR China, but they are most numerous in the more densely populated east of the country. Although the written Mandarin language is uniform, the spoken language differs from province to province with nine major dialects recognised. Hence, it can be argued that use of the term Han to encompass all of these people is more a political convenience than a true measure of ethnicity.

Recognition of the Han as a nationality coincided with the advent of the Republican period in 1908. Dr Sun Yat-sen discerned a need for the Chinese people to develop a sense of nationhood if post-imperial China was to succeed. Sun admired the western nations such as Britain and the United States for their

sense of nationhood and national pride, something that was not present in Chinese society at that time. During this time the figure of the Yellow Emperor (*Huangdi)* was elevated in status to that of a national symbol and declared to be the first ancestor (*shizu*) of the Han *minzu* (Dikötter 1992). The Yellow Emperor was a mythical figure who was thought to have reigned from 2697 BC to 2597 BC. In fact, he became a figurehead for radical nationalist organisations during the first half of the 20th century. This sense of national pride and belonging was necessary in order to mobilise all of the Chinese peoples initially against the rule of the Qing dynasty, and later to establish a strong Chinese republic under a centrally controlled government which in effect was similar to the proceeding imperial dynasties (Gladney 1996).

After the establishment of the Peoples Republic of China in 1949, the Communist Party also used nationalist ideals to exert its power through a centralised government. Almost from the start of the Peoples Republic, there has been continuous evaluation of ethnicity and nationhood (Lipman 1998), and the Han were defined as an official nationality along with 55 other minority nationalities. At first, official recognition of the minority populations fulfilled a promise of ethnic autonomy made during the Long March in 1932 in order to facilitate the survival of the Chinese Communist Party.

Once power had been secured, the PR China government eroded this autonomy for the sake of national unity. For example, the government removed from the constitution a clause allowing ethnic nationalities the right to secede from the Republic, and forced many ethnic peoples into communes with Han Chinese (Gladney 1996). Amongst many other reforms, the 3rd Party Plenum of 1978, restored most of the autonomy of the minorities. The net result was a huge

rise in the numbers of persons recognised as belonging to a minority population, with more and more Chinese citizens identifying themselves as belonging to a specific minority group (Gladney 1996).

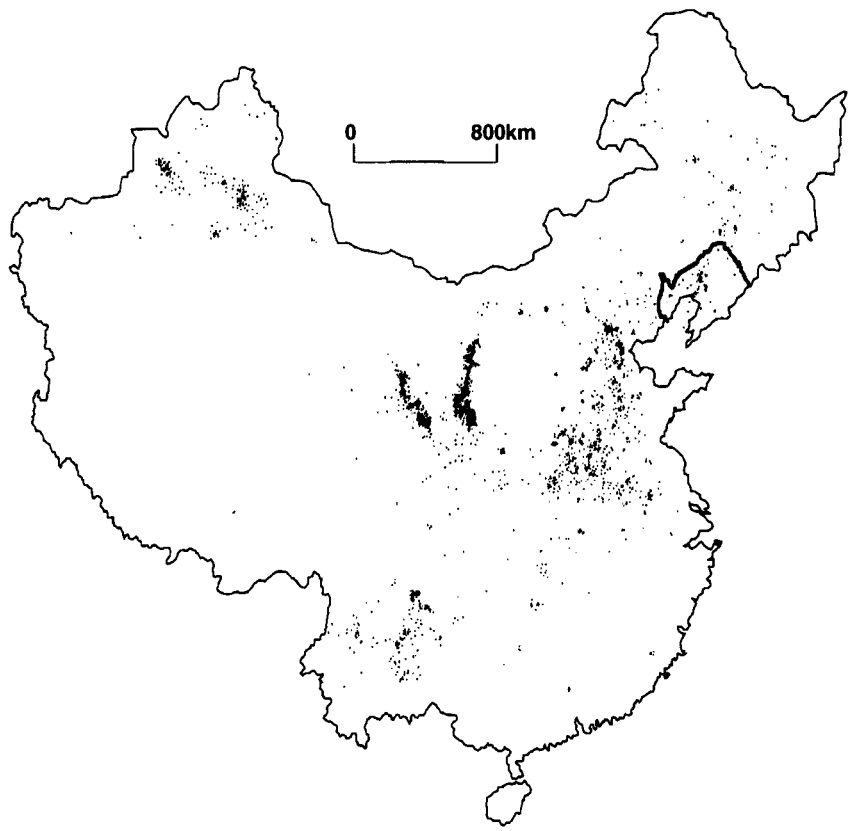
2.6 The Hui

2.6.1 Introduction

According to the 1990 PR China census, the Hui are the largest Muslim community in the country. At 8.6 million, they account for some 50% of the total Muslim population and are resident in 19 of the provinces of PR China (Figure 2.2 (Family Planning Commission 1997). It is notable that the Hui are recognised as an official minority despite their lack of a common territory or a common language. Their extensive geographical spread suggests that the Hui may actually comprise many smaller subpopulations which could claim minority status in their own right, and so the Hui of Liaoning province should perhaps be more properly regarded as a separate distinctive community from other Hui communities.

The origins of the Hui (or Huizu) are diverse, and it is believed that they include individuals whose ancestors originated in pre-Islamic times, from Central Asia, Iran, and the Middle East (Du and Yip 1993). For example, in 61 AD some 10,000 Turkic families from the Central Asian city-states of Samarkand and Bokhara were recorded as settling in the ancient capital of China, Chang'an, now called Xi'an (Du and Yip 1993). Other recorded exoduses to China include the influx of Persian refugees from what is now modern Iran after the fall of the Sassanid Empire in 652 AD.

Figure 2.2 The geographical spread of the Hui minzu over PR China



*Each dot represents 3000 Hui Muslims. Bordered area is Liaoning province. Adapted from Gladney (1998)

2.6.2 The introduction of Islam into China

With the establishment of Arab rule over the Middle East and parts of Central Asia in approximately 652 AD, many Arab rulers made contact with their Chinese counterparts and established trade contacts via the Silk Road. The ancient Silk Road was an overland route that stretched some 7,000 km between Xi'an and Constantinople/Istanbul on the Mediterranean and flourished between 100 BC and 1630 AD (Hopkirk 1980, Yifu *et al.* 1989). During the Tang (618 AD-907 AD), the Five Dynasties (907 AD –960 AD), and especially in the Song period (960 AD- 1279 AD), Arab traders began to actively trade in Chinese cities, mingling with the resident merchants from West and Central Asia. These Arab

traders were largely unaccompanied males who settled in PR China, married Han women, and procreated (Du and Yip 1993, p. 44).

In addition to traders, it is also reported that Arab troops settled in China. In 755AD the Emperor Su Tsung appealed to the second Abbasid Caliph, Abu Ja'far, for help in recapturing Ch'ang-an from a military commander who had rebelled against the Tang emperor and captured his capital city. In response to the Emperors appeal, the Caliph sent 4,000 troops and the city was recaptured. After this successful military campaign, these Arab troops remained in China, intermarried with Chinese and formed a Muslim community (Wong and Dajani 1988).

2.6.3 The Yuan dynasty

It was during the Yuan dynasty (1279 AD –1368 AD) that Muslim ethnic groups came to prominence in China. Muslims were entrusted with vast responsibilities and powers, primarily acting as middlemen between the Mongol overlords and the majority Chinese population (Leslie 1986). The Muslims were employed by the Mongol rulers to administer conquered populations in positions such as Governors, trade commissioners and various other bureaucratic roles. Muslims also were influential in the creation of the Chinese calender, in the building of Peking (now Beijing), in medicine, and in the military (Leslie 1986). In conclusion, during the period of the Mongol reign the Muslims were second in influence only to the Mongols themselves, and they became an important part of the ruling bureaucracy.

2.6.4 The Ming dynasty

With the founding of the Ming dynasty (1368-1662), Islam had been established in China for approximately 700 years (Rahman 1997). Up to this time Muslims had maintained a separate, alien status, with their own customs, language, and traditions. Under the Ming dynasty this situation began to change, as Muslims became fully integrated into Han society (Leslie 1986, Lipman 1998).

Assimilation was strongly encouraged to wipe out the influence of any foreign group and to establish the superiority of Han culture. The practice of male foreigners marrying Han women was specifically encouraged by an Imperial edict issued during the fifth year of the reign of Hongwu which stated that:

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

One result of the integration of Chinese Muslims into Han society was the process by which Muslim names changed. Many Muslim men simply took the name of their Han wife, while others adopted the Chinese character that most closely resembled their own name (Rahman 1997).

2.6.5 The Qing dynasty

During the Ming dynasty, Muslim communities had been almost fully integrated into Chinese society. This changed with the onset of the Qing dynasty (1662 AD – 1908 AD). During the years 1781 - 1784 many new imperial edicts were enacted in which Muslims were lumped together in a single group, with special laws referring to them as such (Leslie 1986). A result of this new imperial policy was a split of the Hui into those who supported the Qing dynasty, mostly communities involved in trade in the eastern regions of the empire, and those who

did not, predominantly resident in the north west of the country (Lipman 1998). The founding of many Islamic sects and philosophies further facilitated this split. The schism on political and religious grounds could have further isolated Hui communities, resulting in an increased restriction on the availability of potential marriage partners and possibly affecting the patterns of genetic diversity of subsequent generations.

Until the seventeenth century, virtually all Muslim communities in China focused their communal life around the local mosque that was run by an *Ahong*, or teacher, appointed by the community elders (Lipman 1998). This local version of Islam is called *Gedimu* and it remains the most predominant style of Islam in China today (Gladney 1996). During the mid- to late Ming, some Muslim communities began to take a greater interest in Islamic history and law, in part as a means of alleviating problems caused by the Ming policy of integration (Leslie 1986). This change coincided with the arrival of Sufism in China. Sufism, sometimes called Islamic Mysticism, revolves around the following of a Sufi or teacher. These Islamic cults first arose in the Islamic heartlands during the 13th century, when the Mongol Empire was at its zenith. Although the orders spread to China at that time, they did not make an impact until the late eighteenth and the nineteenth century (Lipman 1998).

During the Qing period, a number of Sufi orders spread throughout Muslim China, with the teaching of each order based on different interpretation of Islamic texts. The four main Sufi orders founded in China are the Jahriyya, Khufiyya, Qadiriyya and Kubrawiyya. These *menhuan* (saintly lineages) were further subdivided into many smaller *menhuan* and branches created along ideological, political, geographical and historical lines (Gladney 1998).

One result of the growing power of the Sufis was the occurrence of battles between different Muslim orders, and clashes between Muslims and the Imperial army during the eighteenth and especially in the nineteenth century (Gladney 1996, 1998). In one war between the Qing imperial forces and the Sultanate of Dali from 1855-1873 an estimated 10 million persons died (Dessaint 1995-1996). Since the present population of Muslims in China is approximately 16 million (Family Planning Commission 1997), and assuming that around half of the casualties were supporters of the Sultanate, i.e., Chinese Muslims, casualties on this scale may have had a significant bottleneck effect on the present genetic structure of the Hui.

2.6.6. The Republican period

The Republican period (1908 – 1949) followed the fall of the Qing dynasty. An eastern republic was established and, at the same time, there were many different warlord states, some of which were ruled by Muslims. The most powerful of the Muslim rulers was the warlord who controlled Ningxia, Ma Fuxiang (Lipman 1998). He sided with the Guomindang on its ascent to power in 1928, and thus became an important regional leader to the Guomindang leader Chiang Kai-shek. This political power resulted in the establishment of the China Islamic Society, and the reopening of many Muslim schools and other similar religious institutions which had been closed or destroyed during the Qing Dynasty (Lipman 1998).

While the support of warlords was needed by the Guomindang to keep its hold on power, a contrary policy was adopted to the ethnic minorities to suppress ethnic nationalism. In a reference to the Muslim minorities, Chiang Kai-shek stated that the "Mohammedans in present-day China are for the most part actually

members of the Han clan who embraced Islam"(Gladney 1996), and "the differentiation among China's five nations (as first defined by Dr Sun Yat-sen) is due to regional and religious factors, and not to race and blood". Under this policy, the Hui were not considered a separate nation, but a religious group with special characteristics. This policy is still maintained by the government of The Republic of China, Taiwan, which refers to the Hui as a religious group, not an ethnic community.

2.6.7 The Peoples Republic period

As previously noted, the origins of the present system of official minorities in PR China go back to the Long March of 1932, when the Communists were expelled into the regional areas of China by their Guomindang opponents (Fairbank and Reischauer 1990). As a result the Communists had to negotiate with the various minority peoples who lived in western China, but who were wary of this army from the east. Unlike the Guomindang, the Communists promised autonomy to the minorities when they ascended to power, that is recognising ethnic *minzu*, unlike the Guomindang polices. This was confirmed in The Communist Constitution of 1932, which allowed for complete autonomy of various minority regions, prefectures and counties (Gladney 1996).

Almost immediately after the establishment of the Peoples Republic of China, anthropologists and demographers were recruited to survey the population of China. As a result people who were formally called Hui were divided into 10 separate Muslim *minzu*, the Uygur, Kazak, Dongxiang, Kirghiz, Sala, Tadjik, Uzbek, Baonan, Tatar and Hui. The first nine were given separate *minzu* status as they each had a unique language of their own. In essence the modern Hui *minzu*

refer to those Muslims who do not have a language of their own but who speak the dialects of the peoples among whom they live (Gladney 1998).

While autonomous regions, prefectures and counties were set up specifically for these Muslim *minzu*, in reality these communities had limited power and most populations were forced into communes with the effective loss of autonomy (Gladney 1996). Subsequently the PR China government enforced the One China nationalism first espoused by the first Chinese president, Dr Sun Yat-sen and later by Chiang Kai-shek.

The situation was however reversed after the reforms of 1978 when the minority nationalities again gained a bigger voice in China. As a result of these changes, the Hui people have now developed a strong sense of ethnic identity and nationalism.

2.7 Summary

The different concepts of ethnicity indicate that, in defining populations for genetic analysis, great care is needed to differentiate between population groupings which are either a political construct, a scientific construct, or a culturally distinct grouping. The Hui have only recognised themselves as a nationality from the beginning of the 20th Century. Serious arguments about the existence of a Hui *minzu* started to occur in Hui communities with the fall of the Qing Dynasty (Lipman 1998). This change occurred in parallel with the coming to power of various Warlords in North West China, and then the creation of official Hui *minzu* status with the establishment of the Peoples republic in 1949. Before this, the Hui were recognised purely as a religious group, and the term Hui literally translates as Muslim thus encompassing all those who held the Islamic faith and resided in China regardless of their original ethnic origins (Gladney

1996). The Han also are more of a political construct than a culturally distinct grouping, consisting, as they do of different cultural heritages, different language groupings, and even possibly different evolutionary origins (Chu *et al.* 1998). Any genetic analysis of these populations must recognise the complicated underlying social and cultural divisions, and the possible effects on genetic diversity that may be entailed.

Chapter 3

Molecular and population genetics methodology

3.1 Introduction

For the purposes of the present study it was decided that the most appropriate method of assessing genetic variation in the Hui and Han ethnic groups was via analysis based on a panel of microsatellite markers. Since the study is mainly concentrated within a historical rather than an evolutionary time frame, it was hypothesised that the effects of migration, admixture, and isolation would prove to be major factors influencing the differentiation of Hui from Han.

According to historical sources the history of the Hui includes a number of unique features, including male migration, female admixture and consanguinity. Therefore a comparison of autosomal and Y- chromosome microsatellite markers should provide genetic evidence of these influences.

The primary methods used to measure differentiation of microsatellite markers in the Hui and Han populations include a direct comparison of allele frequency distributions, the calculation and comparison of observed and expected heterozygosity, an assessment of linkage disequilibrium, and the calculation of *F*-statistics and *F*-statistic analogues.

3.2 Microsatellite markers

3.2.1 Definition of microsatellite markers

Microsatellites are tandemly repeated sequences whose unit of repetition is between one and six base pairs. Microsatellite sequences, which are randomly dispersed in the genome, are members of a larger group of repetitive DNA sequences and include satellite DNA, minisatellite DNA and transposable elements (Charlesworth *et al.* 1994).

Microsatellites have been proposed as the genetic marker of choice in molecular genetic studies, as they have a high level of polymorphism and are assumed to be selectively neutral according to the definition first proposed by Kimura (1968). That is, while many different alleles may exist, each allele is functionally equivalent. It is assumed that microsatellites play no functional role in the human genome, and so they would have a constant rate of evolution, which is independent of the size of the population.

Microsatellites can be used as markers for the analysis of genetic diversity between populations of the same species because they have very high mutation rate, some tandem repeats have mutation rates as high as 10^{-3} (Goldstein *et al.* 1995, Péréz-Lezuan *et al.* 1997a). As noted earlier, this equates to a correspondingly high level of polymorphism. These features of microsatellites have led to their widespread adoption in paternity testing, linkage analysis, and the reconstruction of human phylogenetic trees (Goldstein *et al.* 1995).

3.2.2 Microsatellite mutation models

The analysis of population structure using microsatellites is dependent on a correct appreciation of the various mutation models applicable to microsatellite markers. Two broad theories pertain, the Infinite Allele Model (IAM), and the Stepwise Mutational Model (SMM).

In basic terms, IAM assumes that every new mutation gives rise to a new allele. Therefore alleles that are identical by state (i.e., the same size) are also identical by descent (i.e., the allele was inherited from a common ancestor). Microsatellite mutation does not fit this model exactly. Initially it was believed that microsatellite mutation involved the gain or loss of a single repeat unit (Weber and Wong 1993), although more recently some microsatellites have been found to mutate by more than one repeat unit at a time (Di Rienzo *et al.* 1994). Empirical observation has further demonstrated that through the course of several generations microsatellites can, for example, lose a repeat and a generation later regain that repeat. This phenomenon, termed homoplasy, results in alleles that are identical by state but not necessarily identical by descent. Since microsatellite mutation rates are high, levels of homoplasy among microsatellite alleles are assumed to be equivalently large (Goldstein *et al.* 1995).

The SMM model, first defined by Kimura (1968), can be utilised in microsatellite nucleotide sequences to simulate single step mutation, with the possibility of high rates of homoplasy (Goldstein *et al.* 1995, Takezaki and Nei 1996), and parameters such as the Ddm (delta mu squared) genetic distance is based on this theory. Ddm is based on the square of the difference in the means of allele size between populations, with a comparison of allele size difference acting as the main parameter for SMM distances.

Since the formulation of the various mutation theories, there has been widespread debate as to which distance measure is best suited for use with microsatellite data. The very structure of microsatellites is complicated. They can comprise repeats ranging from 1-6 base pairs long, and have one of three structures, pure, compound and interrupted (see Table 3.1)

Table 3.1. Microsatellite classification

Pure	CACACACACACACACACACACACA
Compound	CACACACACACACACACACAGAGAGA
Interrupted	CACATTCACACACACACATCACATCACA

This complexity results in different types of microsatellite that mutate in different ways and at different mutation rates. A study by Shriver (1993) attempted to compare the observed genetic diversities of various classes of Variable Number of Tandem Repeat loci (VNTRs), with simulations derived from the SMM model. The sequences studied included di-, tri-, tetra- and pentanucleotide microsatellites, and minisatellites composed of 15-70 base repeats. The results indicated that all microsatellites with 3-5 base repeats, 65% of microsatellites with dinucleotide repeats, and 27% of minisatellites matched the corresponding simulation values. It was concluded that minisatellites, and to a lesser degree, dinucleotide microsatellites, are more similar to the expectations of the IAM model than to the SMM model. In theory, it may be possible to yield information on evolutionary origin and on recent genetic drift from the same microsatellite data, depending on the type of microsatellite and the mutation model that is utilised.

3.2.3 The human Y-chromosome and microsatellites

The strong pattern of male migration, as suggested by historical information on the Hui, indicates another possible site for the study of genetic variation, i.e. the Y-chromosome. The first Y-chromosome polymorphisms were reported more than a decade ago (Casanova *et al.* 1985, Lucotte and Ngo 1985), however progress in elucidating further examples has been slow because conventional DNA polymorphisms have been difficult to find on the Y-chromosome, and those that have been discovered often have proved to be of limited informativity (Jobling and Tyler Smith 1995).

Recently, a series of polymorphic microsatellites have been developed and tested on many different population samples from around the world (Kayser *et al.* 1997, de Knijff *et al* 1997). These microsatellites (DYS19, DYS388, DYS389I+II, DYS390, DYS391, DYS392 and DYS393) are highly polymorphic, compared to other Y-chromosome polymorphisms, but are still less polymorphic than their autosomal counterparts (Péréz-Lezuan *et al.* 1997b). Nonetheless, they have proven useful as tools for forensic analysis, in such applications as stain analysis and paternity analysis (Kayser *et al* 1997).

Y-chromosome microsatellite markers have also proven useful in population genetics because of the ability to perform accurate haplotype analysis. By their nature, Y-chromosomes are effectively haploid, with only a small portion of the chromosome that can undergo limited recombination with the X-chromosome. The lack of recombination means genetic diversity is more limited in Y-chromosomes than in autosomal chromosomes, but in turn, the absence of recombination increases the effect of genetic drift. It is this property of Y-chromosomes that could prove very useful in elucidating genetic differences

between closely related populations whose time of divergence has been relatively short (de Knijff *et al* 1997).

The advent of these probes and other Y-chromosome markers subsequently developed, has allowed geneticists to exploit the haploid nature of Y-chromosomes, to provide a unique insight into human genetic variation via the construction and analysis of Y-chromosome haplotypes (Cooper *et al.* 1996, Roewer *et al.*1996). One such study showed that the Finnish people have both European and Asiatic origins (Kittles *et al.* 1998). Other analyses have indicated specific patterns of geographic clustering, allowing scientists to pinpoint the origins of human populations through male gene flow (Malspina *et al.* 1998, Zerjal *et al.* 1997).

3.3 Population genetics methodology

3.3.1 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium is a central facet of population genetics theory (Hartl and Clark 1997 p 74). As the name suggests, the concept was first defined by Godfrey Hardy, an English mathematician, and Wilhelm Weinberg, a German physician. Through mathematical modelling, it was concluded that gene pool frequencies are inherently stable but that evolution should be expected in all populations virtually all of the time. This apparent paradox was resolved by analysing the probable net effects of evolutionary mechanisms. Hardy, Weinberg, and the population geneticists who followed them came to understand that evolution would not occur in a population if seven preconditions were met:

Table 3.2 Seven conditions for the non-occurrence of evolution in a population

1. Mutation is not occurring
2. Natural selection is not occurring
3. The population is infinitely large
4. All members of the population breed
5. All mating is totally random
6. Everyone produces the same number of offspring
7. There is no migration in or out of the population

In other words, if no mechanisms that can cause evolution to occur are acting on a population, evolution will not occur and the gene pool frequencies will remain unchanged. However, since it is highly unlikely that any of these seven preconditions, let alone all of them, apply in the real world, evolution is the inevitable result.

A simple equation was developed that can be used to determine the genotype frequencies in a population and to track their changes from one generation to another. This has become known as the "Hardy-Weinberg equilibrium equation". This equation is defined in terms of a biallelic locus where p is defined as the frequency of the first allele and q as the second allele for a locus consisting of a pair of alleles (A and a). This final equation is as follows:

$$p^2 + 2pq + q^2 = 1$$

Equation 3.1 Hardy-Weinberg equilibrium for a bialleic locus

In this equation, p^2 is the frequency of homozygous (AA) individuals in a population, 2pq is the frequency of heterozygous (Aa) individuals, and q2 is the frequency of those who are homozygous (aa). Deviation from equilibrium occurs when the observed frequency is significantly different from that predicted by the above equation.

Microsatellite loci are polymorphic and therefore the two-allele model could not apply. A more general equation is shown below where p_I is the allelic frequency of A_I and p_j is the allelic frequency of A_j.

$$\sum_i p_i^2 A_i A_i + \sum_{i,j} 2 p_i p_j A_i A_j$$

Equation 3.2 Hardy-Weinberg equilibrium genotypic array

Detection of deviation from Hardy-Weinberg equilibrium is usually tested for statistical significance. Traditionally this was accomplished using χ^2 "goodness of fit tests" (Guo and Thompson 1992). Testing relies on asymptotic results, which may or may not be a characteristic of the data, making the level of statistical confidence low. The alternative is computation of Fisher's exact tests. In the past, due to a lack of computing power, use of the exact test has been restricted to biallelic loci assessed from small sample sizes. With the development of appropriate computational power, these calculations are now commonplace and they have all but superseded goodness of fit methods.

Even with the statistical methods available, by direct interpretation of the theory, some deviation from Hardy-Weinberg equilibrium (HWE) might be expected in all human populations, as they could not meet all of the seven criteria listed in Table 3.2. However, the interaction of mutation, selection, migration, admixture and inbreeding, which happens to various degrees in every human

population, can produce an effect tantamount to no statistically significant deviation being detected (Guo and Thompson 1992). But if one effect is clearly the dominant factor over the others in a particular population, deviations may be observed.

3.3.2 Linkage equilibrium

Associations between alleles can be expanded from associations of alleles within one locus, to the association of alleles at two different loci. This random gametic association between two alleles from two different loci is called linkage equilibrium. It occurs with random mating and any deviation from this is called linkage disequilibrium.

Linkage disequilibrium is not necessarily correlated with linkage, as alleles at different loci may have frequencies that show association whether or not they are linked (Weir 1996). In other words, loci from two separate chromosomes can show associations.

One commonly used measure of linkage disequilibrium is the linkage disequilibrium parameter (D). This is most easily defined for two autosomal biallelic loci. Two loci, A and B each have two alleles A_1, A_2 (at frequencies p_1 *and* q_1), and B_1, B_2 (at frequencies p_2 and q_2) respectively. Given $P_{11} = p_1q_1$, $P_{12}=p_1q_2$, $P_{21} = p_2q_1$ and $P_{22} = p_2q_2$ the linkage disequilibrium parameter is:

$$D = P_{11}P_{22} - P_{12}P_{21}$$

Equation 3.3 Linkage disequilibrium parameter

The parameter D is therefore defined as the difference between the observed frequency of a gametic type, and the frequency expected on the basis of

random association of alleles (Slatkin 1994). Larger values of D suggest increased levels of linkage disequilibrium in a population.

A second, more commonly used definition of linkage disequilibrium, as with HWE, is in the sense of statistical difference. In this usage, linkage disequilibrium exists between two loci if a statistical test shows there is a significant non-random association between any two alleles at the respective loci. The most commonly employed statistical method, as used to assess deviation from HWE, is Fisher's exact test (Slatkin 1994). However these tests can be computationally difficult. Therefore it is only since the development of rapid algorithms for performing Fisher's test using Monte Carlo methods to approximate the results from an exact test that these analyses have been possible. Population genetics software, such as GENEPOP (Rousset 1995) and ARLEQUIN (Excoffier *et al.* 1992), include algorithm methods for the computation of HWE and linkage disequilibrium

While HWE tests internal population structure, linkage disequilibrium measurement has been primarily used as a gene-mapping tool, usually as an alternative to linkage analysis when the number of informative families for a trait is exhausted. More recently it has been proposed as a method for the genetic analysis of the demographic history of populations, e.g., population growth and decline, mating structure and migration and isolation (Slatkin 1994, Weiss 1998). Simulation results (Slatkin 1994), and surveys of genome databases (Peterson 1995 *et al.*), have shown that the extent of statistically significant disequilibrium depends both on the recombination rate between loci and on the demographic history of the population from which the samples were obtained.

3.3.4. Consanguinity

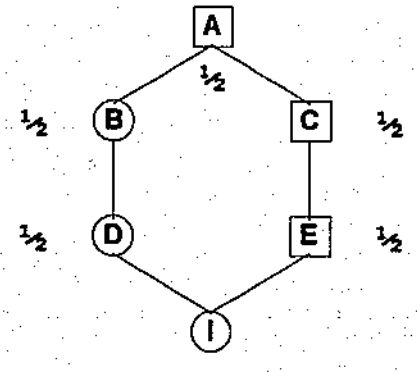
Inbreeding can be a major cause of deviation from HWE, and possibly linkage equilibrium. The basic observation of inbreeding is that of mating between biological relatives. Two individuals are said to be related if among the ancestors of the first individual are one or more ancestors of the second individual. Because of shared common ancestors, these two individuals could share genes at one or more loci that are identical. Identical gene copies that are due to shared ancestry are said to be identical by descent (IBD).

Identity by descent in human populations is defined in terms of consanguinity. In communities with a tradition of close kin unions, marriages between persons biologically related as second cousins or greater are generally categorised as consanguineous (Bittles 1998). However, in populations where consanguinity generally is avoided or proscribed, the definition may be extended to cover third cousin or more remote unions. In assessing the degree of consanguinity in a pedigree, the most commonly used measure is the pedigree inbreeding coefficient.

F represents the probability that the offspring is homozygous due to identity by descent at a randomly chosen autosomal locus. As such, F is a probability that can range in value only from zero (no loci identical by descent) to one (all loci identical by descent). F therefore can be calculated for an individual if the pedigree of the individual is available and it shows a marriage(s) between biological relatives in previous generations. For example, in Figure 3.1, I is the child of a first cousin marriage between D and E with a common ancestor A. F is calculated by tracing the paths of the gametes that lead from the parents of I back to A, through B and C. The probability of the alleles being identical by descent is

one-half because, with Mendelian segregation, the probability that a particular allele present in a parent is transmitted to a child is one-half (Hartl and Clark 1997 p149-152).

Figure 3.1 Abbreviated diagram of a first cousin marriage.



Therefore, in the present example, there are five paths between I and A, and so the probability of identity by descent is 1/2 * 1/2 * 1/2 * 1/2 * 1/2 or 1/32. This can be simplified to the equation:

$$F = \sum \left(\frac{1}{2}\right)^{n_1+n_2+1}$$

Equation 3.4 Pedigree inbreeding coefficient.

where n is the number of individuals separating the child and the common ancestor and 1/2 is the probability that the child will inherit the allele of a specific parent. The child of a first cousin marriage will have two such paths, one for each grandparent and so the probability of autozygosity is 1/32 + 1/32 = 1/16. Hence the child of a first cousin union is predicted to display 6.25% greater homozygosity than the child of a non-consanguineous marriage.

In a large pedigree with a history of consanguinity prior to the current generation, the actual coefficient of inbreeding predicably will be higher than can be calculated for a single generation, due to the cumulative effect of inbreeding (Shami *et al.*, 1994). The effect of prior inbreeding is calculated as the sum of the probability of identity by descent due to each separate path of inheritance of the alleles, and is represented by the equation:

$$F = \sum \left(\frac{1}{2}\right)^{n} (1 + F_A)$$

Equation 3.5 Pedigree inbreeding coefficient: multiple generations.

where n is the number of individuals in each path connecting the parents and A is the common ancestor in each path (Hartl and Clark 1997 p 149-152).

3.3.5 Inbreeding as a deviation from random mating expectations

Another definition of inbreeding is in terms of a system of mating definition at the population level. It is useful in this context to examine deviations from the genotype frequencies expected in HWE that are due to inbreeding. It can be shown that with inbreeding the allele frequencies remain the same, only genotype frequencies change (Hartl and Clark 1997 p126).

The work of Wright (1951) determined that the deviation from random mating expectations could be calculated by what is known as the covariance, COV, between uniting gametes. With random mating, COV= 0 but with non-random mating, this COV can be either positive or negative in sign.

The actual correlation between uniting gametes is COV/pq, so one way to define the correlation coefficient is:

$$f = \frac{COV}{pq}$$

Equation 3.6 Correlation coefficient

Defined in this way, as shown in Table 3.3, genotype frequencies with inbreeding can be expressed as:

Table 3.3 Deviation of genotype frequencies from Hardy-Weinberg expectations

Genotype	AA	Aa	aa
Frequency	p^2+pqf	$2pq(1-f)$	q^2+pqf

Note that a positive correlation between uniting gametes leads to a heterozygote deficiency in the population, and a negative correlation gives an excess of heterozygotes. Hence, f measures deviation from Hardy-Weinberg genotype frequencies but not the probability of being identical by descent. The value of f can range from –1 to 1 and denote inbreeding ($f>0$), random mating ($f=0$) and avoidance of inbreeding ($f<0$). By comparison, avoidance of inbreeding cannot be measured by F.

3.3.6 The inbreeding effect of population subdivision

Another definition of inbreeding is the inbreeding-like effect that occurs due to population subdivision. Wright (1951) found that the effect of population subdivision could be measured by a quantity called the fixation index, F_{ST}. Given that p_1 and q_1 are the frequencies of alleles A and a in one population, and p_2 and q_2 are the frequencies of alleles A and a in a second population then:

$$F_{ST} = \frac{Var(p)}{\overline{pq}}$$

Equation 3.7 Wrights F_{ST}

where $\overline{p} = \frac{1}{2}(p_1+p_2)$, $\overline{q} = \frac{1}{2}(q_1+q_2)$ and $\mathrm{Var}(p) = \overline{p^2} - \overline{p}^2$. Wright's F_{ST} can be put into context by examining a pooled population composed of two populations. The pooled population genotypes can be described using parameters defined in equation 3.7 as follows:

Table 3.4 Deviation of pooled genotype frequencies from Hardy-Weinberg expectations

Genotype	AA	Aa	aa
Frequency	$\overline{p}^2 + pqF_{ST}$	$2\overline{pq}(1-F_{ST})$	$\overline{q}^2 + \overline{pq}F_{ST}$

Table 3.4 shows that subdivision of the population into genetically distinct subpopulations causes deviations from HWE that are identical in form to those caused by the inbreeding system of mating within a population shown in Table 3.3. Wright went further and described a relationship between deviations from HWE within a population to deviations from HWE between populations as follows:

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

Equation 3.8 Relationship of F_{IT}, F_{ST} and F_{IT}

where F_{IS} is the inbreeding coefficient due to the reduction in heterozygosity of an individual due to non-random mating within a population, and F_{IT} is the overall inbreeding coefficient of an individual.

The most widely used interpretation of these statistics is that determined by Nei (1973), who defined the statistics for multiallelic multilocus information from any number of subpopulations as follows:

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

Equation 3.9 Heterozygosity of an individual in a subpopulation

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

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

Equation 3.10 Heterozygosity of an individual population

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

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

Equation 3.11 The expected heterozygosity of an individual in an equivalent random mating total population.

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

(a) $$F_{IS} = \frac{\overline{H}_S - H_I}{\overline{H}_S}$$ *(b)* $$F_{ST} = \frac{H_T - \overline{H}_S}{H_T}$$ *(c)* $$F_{IT} = \frac{H_T - H_I}{H_T}$$

Equation 3.12 (a) – (c) Nei's interpretation of Wright's F-statistics

where F_{IS} is defined by Nei as the reduction in heterozygosity of an individual due to non-random mating within a subpopulation, F_{ST} is the reduction in heterozygosity of a subpopulation due to random genetic drift, and F_{IT} is the reduction in heterozygosity of an individual relative to the total population.

In addition to manipulating Wright's original statistics, Nei developed his own F-statistic analogues, called G_{ST} statistics. These statistics were based solely on the partitioning of gene diversity rather than the inbreeding effect on heterozygosity. Gene diversity, while computationally the same as expected heterozygosity, is defined as the probability of two randomly chosen genes from the same population being different.

$$D = 1-\Sigma x_i^2$$

Equation 3.13 Nei's gene diversity

where x equals the population frequency of a locus in the ith population. An estimate of this population gene diversity can be calculated from the sample gene diversity. These formulae for gene diversity can be extended for population subdivision as follows:

$$H_S = 1-\Sigma x^2_{ki}$$

Equation 3.14 Gene diversity within a subpopulation

where x_{ki} is the frequency of the ith allele in the kth subpopulation.

$$D_{ST} = \sum_k \sum_l [[\Sigma_i (x_{ki} - x_{li})^2 / 2] / s^2]$$

Equation 3.15 Coefficient of gene differentiation

where x_{li} is the frequency of the ith allele in the lth locus and s is the number of subpopulations. It then follows that:

$$H_T = H_S + D_{ST}$$

Equation 3.16 Gene diversity in the total population

and

$$G_{ST} = D_{ST} / H_T$$

Equation 3.17 Average gene diversity between subpopulations

The advantage of this method is that haploid systems and other systems not obeying HWE can be analysed using these statistics, as only gene diversity is utilised and not heterozygosity.

Earlier, Cockerham (1969) developed a convenient way to position F-statistics into a familiar context for hypothesis testing. This involves the partitioning of F by analysis of variance. The result was a set of formulae solved to give F-statistics in terms of variance components.

$$\sigma^2_T = \sigma^2_{A+}\sigma^2_B + \sigma^2_W$$

Equation 3.18 Total variance of allele frequency

That is σ^2_T, the total variance of allele frequency within a population, is equal to the sum of its components. These components are, σ^2_A, between subpopulation variance in allele frequency; σ^2_B,between individuals within subpopulation variance in allele frequency; and σ^2_W, between gametes within individuals variance in allele frequency.

Using this approach Cockerham defined three F-statistic type parameters. F, the correlation of genes within individuals , θ, the correlation of genes of

different individuals in different populations; and *f*, the correlation of genes within individuals within populations. This resulted in the formula:

$$f = (F-\theta)/(1-\theta)$$

Equation 3.19 Cockerham's correlation of genes

These parameters can be defined as ratios of the various components of variance as:

$$(a)\quad F = F_{IT} = \frac{\sigma_A^2 + \sigma_B^2}{\sigma_T^2} \quad (b)\quad \theta = F_{ST} = \frac{\sigma_A^2}{\sigma_T^2} \quad (c)\quad f = F_{IS} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

Equation 3.20(a)-(c) Cockerham's F-statistics

By relating *F*-statistics to analysis of variance, it is possible to add levels of population subdivision, and to extract their effects as additional components of variation.

Weir and Cockerham (1984) further developed this concept for a multiallelic multilocus hierarchical structure. Their model involves the weighted sum of variances. Mathematically, the method is much more complex than the definition by Wright or Nei, but leads to unbiased estimators that are statistically more robust. One example is an F_{ST} analogue called $\theta_{\overline{w}}$ which can be estimated from:

$$\theta_{\overline{w}} = \frac{\sum_i \sum_u \sigma_B^2}{\sum_i \sum_u \sigma_T^2}$$

Equation 3.21 Weir and Cockerham's θ_w

where the variances in allele frequency are summed over all alleles *i* and all loci *u*. The advantage of an analysis of variance approach is that the proportion of variances are additive, making possible a wide range of hierarchical population structures.

With the advent of large-scale analyses of mtDNA sequences, a haploid sequence approach to population subdivision was necessary as the previous methods involved departures of allele frequencies from panmictic genotypic expectations. HWE predictions were not applicable to mtDNA sequence data, being haploid. An approach using phenetic and evolutionary distances between haplotypes was developed for mtDNA by Excoffier *et al.* (1992). The method is termed Analysis of Molecular Variance (AMOVA).

The variance components produced from AMOVA analysis can be derived to produce an analogue to *F*-statistics called Φ statistics (Excoffier 1991). These Φ statistics reflect the correlation of haplotypic diversity at different levels of hierarchical subdivision. AMOVA offers the advantage of an analysis of variance model without the requirement for many of the assumptions, such as the assumption of normality. This is achieved through the utilisation of non-parametric permutation tests to assess statistical significance instead of using the asymptotic approach adopted in ANOVA.

The technique is adequate for haploid sequence systems such as mtDNA, but its application to diploid data is much more difficult, especially if gametic phase is unknown. The problem was solved by Michalaklis and Excoffier (1996) who defined AMOVA for a diploid system, with Φ_{ST} being analogous to θ_{π} using maximum likelihood techniques. The result was a population subdivision

technique applicable to all types of genetic data, based on the partition of within population and among population variance of difference in the number of alleles of each haplotype. This enables different genetic data types to be compared (i.e., allozymes, RFLPs, microsatellites, mtDNA sequences), as well as specific formulae for population subdivision analysis using microsatellites (Michalakis and Excoffier 1996).

All of the F-statistic models presented so far share one common assumption, the assumption of the Infinite Allele Model. As previously discussed, microsatellite mutation mostly follows the SMM. Therefore, the previous F-statistic methods based on the IAM may not be appropriate. An SMM alternative to F_{ST} was defined by Slatkin (1995) and called R_{ST}:

$$R_{ST} = \frac{\overline{S} - S_w}{\overline{S}}$$

Equation 3.22 Slatkin's R_{ST}

where R_{ST} is essentially a F_{ST} analogue for microsatellite data that takes into account the difference between microsatellite allele sizes, where S_W is the average squared difference in allele size between pairs of genes within populations and S is the average square size between pairs of genes taken from a collection of populations.

Rousset (1996) further defined this approach with the formulation of the analogous Rho statistics, a SMM equivalent of F_{IS}, F_{ST} and F_{IT}, labelled ρ_{IS}, ρ_{ST}, and ρ_{IT}. The parameter ρ_{ST} is related to Slatkin's R_{ST} in the following manner:

$$R_{ST} = \frac{(1-c)\rho_{ST}}{1-c\rho_{ST}}$$

Equation 3.23 Correlation between Rousset's ρ_{ST} and Slatkin's R_{ST}.

where $c = (2s_s - 1)/(2s_s n_s - 1)$, s_s is the sample size, and n_s is the sample number.

Michalakis and Excoffier (1996) also defined their Φ_{ST} with regard to the *R*st estimate ρ_{ST}. This variant of AMOVA is modelled for differences in allele size and not, as in the original, differences in the number of alleles. Thus, for microsatellite data AMOVA can be performed in two modes, according to differences between haplotypes due to the number of alleles not shared by both haplotypes summed over all loci, and according to differences between haplotypes due to the squared difference in allele size summed over all loci.

3.3.5 Genetic distance

Genetic distance has been defined as the extent of gene difference between populations or species that is measured by some numerical quantity (Nei 1987). Nei (1987) classified two types of distances, the first of which included measurements for population classification whereas the second was applicable to evolutionary study.

Distances that are appropriate for contemporary population comparisons are applicable for population classification. Distances such as D_S (Nei 1973), and D_{ps} (Bowcock *et al.* 1994), are two of the measures that have been used. Nei's standard distance, D_S is the most commonly used genetic distance and is defined as follows:

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

Equation 3.24 Nei's standard genetic distance.

where *Jx* and *Jy* are the average homozygosities across loci in populations X and Y.

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

Distances based on the SMM model make use of the difference in size between alleles (Goldstein *et al.* 1995). An example of this distance measure is Ddm, or delta mu squared.

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

Equation 3.25 Goldstein's delta mu squared genetic distance

where μ_A and μ_B are the means of allele size, summed over all loci, in populations A and B respectively. These distance measures are quite accurate in measuring evolutionary genetic distance as they maintain a linear relationship to time over a long period, for example, several thousand generations.

3.4 Summary

To date, microsatellites provide the most approachable and informative method of analysing human genetic diversity. They are among the easiest and most polymorphic markers to detect and analyse and, as microsatellites are PCR typable only minute DNA samples are needed for allele detection. Microsatellite markers are also the most polymorphic markers available, allowing the detection of variation between closely related populations.

Information from microsatellite analysis involves the use of various statistical and genetic methods to define population structure and genetic diversity. As previously discussed, most microsatellite mutations involve the addition or subtraction of a small number of repeat units a process called the Stepwise Mutation Model (SMM) (Kimura 1968). To account for these observations, various genetic parameters have been proposed for microsatellites, for example, Slatkin's R_{ST}. The advantage of such an approach for this study is the fact that SMM-based parameters are not subject to sample size bias (Goldstein *et al.* 1995). This is important in the present study, as there is a substantial difference in the sample sizes collected from the Han and Hui of Liaoning.

However, the study is focused on genetic diversity developed in a human historical timeframe. Given that most dinucleotide microsatellites mutate at around $10^{-3} - 10^{-4}$ per generation, mutation would not be expected to exert a major effect on genetic diversity in the Han and Hui in the timeframe studied. Furthermore, it was concluded by Péréz-Lezuan *et al.* (1997b) that genetic drift, not mutation, plays the main role in generating the microsatellite variation which has been observed among human populations.

Consequently, it was decided that the combined use of both SMM and non-SMM based statistics would be the best approach to inferring relationships among the Han and Hui. In conclusion, only by the comparison of several different measures and methods with other types of evidence, such as historical information, can the genetic relationship between two populations be inferred with any confidence.

Chapter 4

Sample collection and experimental methodology

4.1 Sample collection

Dr Wei Wang and his Chinese colleagues collected all blood samples on to 3MM Whatman® filter paper. The acquisition of blood spots from Hui community members in Liaoyang, Liaoning province, was administered through the permission of local religious leaders. Religious leaders who were interested in the proposed study organised finger prick blood spot collection from community members on site in their villages. The two pedigrees in the study were obtained via the permission and involvement of the elders of each family. Each person who consented to give blood did so after signing a consent form.

The Han samples were obtained on a random basis from volunteers who provided finger prick blood spots at the Peoples Liberation Army (P.L.A.) No 201 Hospital Liaoyang, P.R. China. A signed consent form also was a prerequisite for the acquisition of blood samples from these volunteers. All individuals sampled were from various co-resident communities in the city Liaoyang, located in central Liaoning province, PR China.

4.2 Extraction of DNA

The collected blood spots were forwarded by courier to the Centre for Human Genetics at ECU for storage at -80^0 C. In this study, 102 random Han, all male, and 53 random Hui samples, from 27 males and 26 females, were analysed, as well as samples obtained from two Hui kindreds numbering 31 and 14 individuals.

For each individual DNA was isolated from two blood spots, using proteinase K treatment followed by phenol/chloroform extraction and isopropanol precipitation at –20^{0}C. The blood spots were cut from the filter paper, quartered, and placed in a 1.5ml microtube with 250 µl 0.1% Triton X-100 and 15 µl

10mg/ml proteinase K at room temperature. The sample was mixed gently for 1 minute, before incubation on a heating block at 50^0C for 30 minutes. This step was repeated once. At the completion of the second incubation period, 25 µl 10x SET buffer (500mM Tris pH 8.0, 50mM EDTA, 5% SDS) was added to the sample and mixed. 500 µl of 1:1 chloroform/phenol was then added and mixed by inversion for at least 10 minutes. The sample was centrifuged for 30 minutes at 13,000 rpm, and the supernatant transferred to a fresh microtube with waste paper materials excluded. A 1/10 volume of 3M Na acetate pH 4.9 and 1 volume of 100% isopropanol were added to the supernatant, mixed and the tube was left overnight at -20^0C to precipitate the DNA from solution. The sample tube was then centrifuged for 30 minutes at 13,000 rpm. The supernatant was discarded, and the DNA pellet washed with ice cold 70% ethanol by inversion. The ethanol was then removed, after another 10 minute centrifugation at 13000 rpm, and the microtube left to dry at room temperature for an hour. Finally, the pellet was resuspended in 25µl autoclaved distilled water.

4.3 Measurement of DNA concentration via spectrophotometry

The DNA concentration and sample purity of the samples was analysed with a Beckman DU 640 UV spectrophotometer. This was achieved by measuring the optical density of a twenty-fold dilution of the original DNA solution at wavelengths of 260 nm and 280 nm. Measurement at 260 nm detects nucleic acids while 280 nm detect proteins. A ratio of 1.6:1 or above indicates satisfactory purity of nucleic acids for amplification. The spectrophotometer was blanked with dH2O used in the dilution of the DNA solution. A concentration of at least 5-10ng/µl of DNA was required for successful PCR amplification.

4.4 Microsatellite markers

The autosomal markers analysed in this study were chosen from a panel of markers recommended by Stanford University (see figure 4.1). Y-chromosome markers used in this study were chosen from a panel of markers recommended by the Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University (Appendix A). Linkage and cytogenetic maps of chromosomes 13 and 15 indicating the positions of the markers used in the study are shown in figure 4.1. As the Y-chromosome is effectively haploid, only a cytogenetic map could be created. Most of the Y-chromosome markers used in the study have yet to be located to a particular position on the chromosome, as seen in figure 4.2.

Figure 4.1 Comparative linkage maps locating microsatellite markers analysed from chromosomes 13 and 15. The distances are measured in centiMorgans.

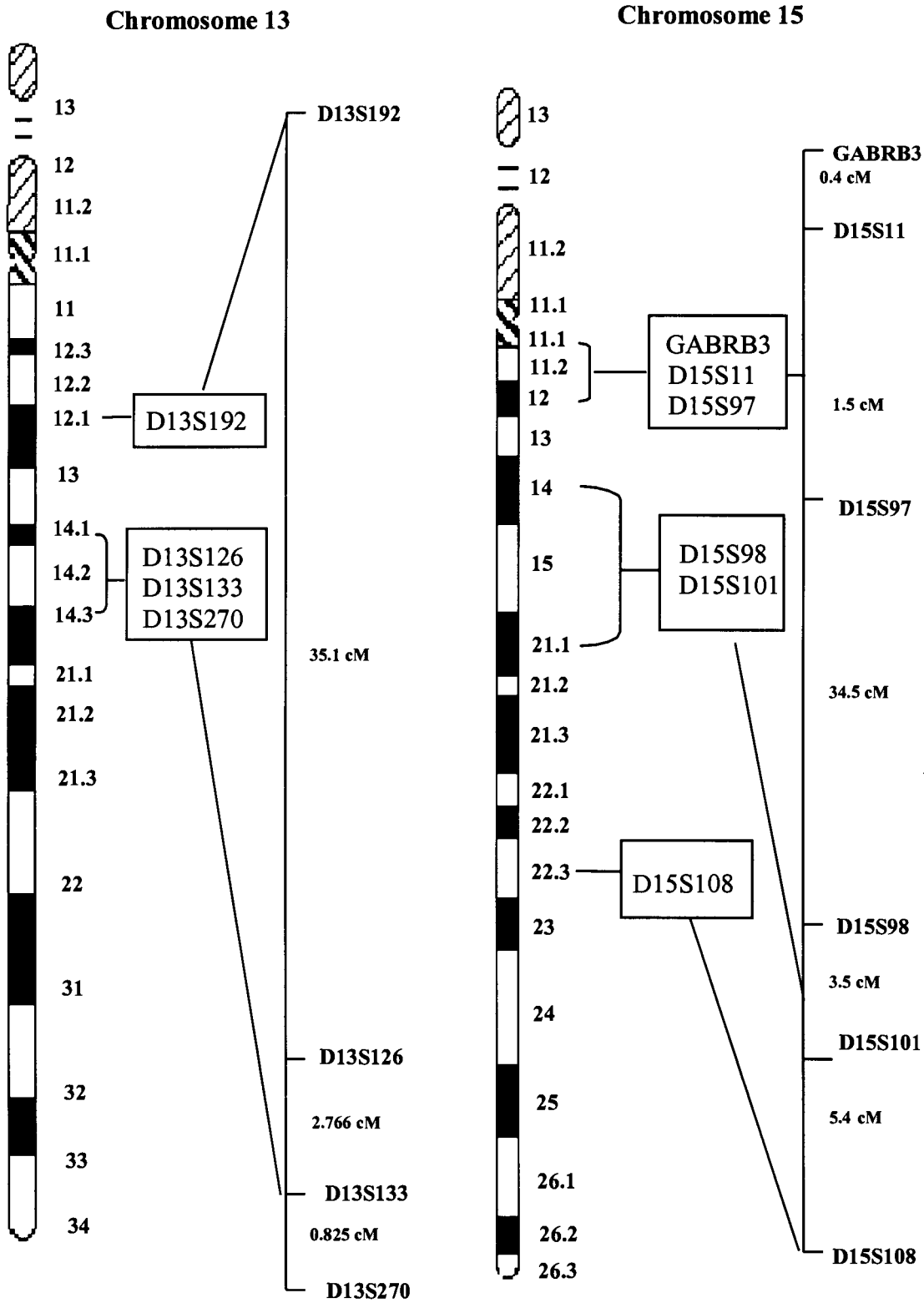
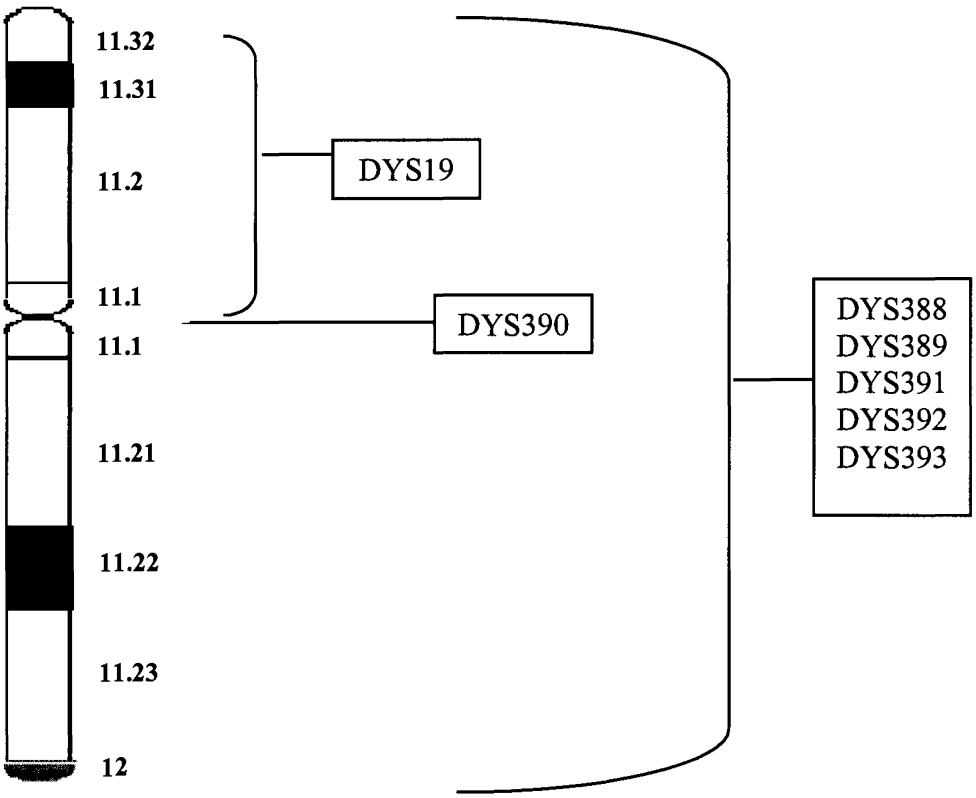


Figure 4.2 Cytogenetic map of the Y-chromosome indicating the approximate positions of the microsatellite markers used.



4.5 Polymerase chain reaction

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

4.6 PCR protocol for autosomal markers

Each autosomal PCR was made up to 5µl, containing 2µl (5-10ng/µl) of target DNA, 0.5µl of forward primer and 0.5µl of reverse primer, 1µl of 5x buffer (1.5mM $MgCl_2$ solution [Perkin Elmer]; 1mMdNTPs, 0.5µl 10xpolymerase buffer (Perkin Elmer), 0.05µl of Amplitaq Taq polymerase (Perkin Elmer®) and 0.95µl of dH2O. Due to variability in the quality of the target DNA and the fidelity of the primers, $MgCl_2$ concentration was varied between 1.5mM and 3.5mM. D13S126, D13S133, D13S192, D13S270, D15S11, D15S101, D15S108 and GABRB3 were successfully amplified by a touchdown procedure which consisted of four main components. Initially, the samples were denatured for 5 minutes at 94°C. This was followed by fifteen denaturing cycles of 20 seconds at 94°C, one minute of annealing, starting at 63°C and reducing in each cycle by 0.5°C (giving a final temperature of 55.5°C), and a 30 second extension period at 72°C. A

further fifteen cycles followed, each consisting of 20 seconds denaturing at 94°C, 20 seconds annealing at 55°C, and 30 seconds of extension at 72°C. The cycle concluded with a five-minute extension period at 72°C.

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

4.7 PCR protocol for Y-chromosome markers.

Each Y-chromosome PCR mixture was made up to 10µl, containing 5µl of target DNA solution, 1µl of 10xbuffer solution (containing 1.5mM of $MgCl_2$ (QIAGEN®), 0.04 µl of 25mM $MgCl_2$ solution, 0.20µl of 10mM dNTPs, 0.50µl of forward and reverse primer containing 100ng of primer oligonucleotides, 0.05µl of Hot Star™ Taq polymerase (QIAGEN®) and 3.21 µl of distilled water.

Qiagen® Hot Star™ Taq polymerase was used for Y-chromosome markers as experimentation with this enzyme produced greater amplification and fidelity of the markers, most of which proved difficult to amplify. The microsatellite marker DYS391 consistently proved very difficult to amplify from a majority of Han and Hui samples, and consequently its use was discontinued.

4.8 Agarose gel electrophoresis

To ensure that the microsatellite sequences were amplified, PCR products were tested by agarose gel electrophoresis. This technique employs an electric current to move the negatively charged DNA towards a positively charged electrode through an agarose gel. The larger the allele fragment, the slower it moves through the agarose. Therefore the alleles are differentiated by length

fractionation. The PCR products are visualised using ethidium bromide (EtBr) under fluorescent light.

A 3% agarose gel solution (3g agarose powder, Sigma Chemical Company) in 100ml 1xTAE buffer (0.04M Tris-acetate; 0.001M EDTA) was prepared and poured on to an 8.5cm x11cm mini-gel tray, and a small toothed comb was inserted at one end of the tray. The gel was allowed to set for approximately forty minutes at room temperature, after which the comb was removed and the gel was then placed in the electrophoresis unit. 3µl of the PCR products were loaded into the wells with 3µl of 6x Fico loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF; 15% Ficoll (Type 400; Pharmacia)). pUC19 DNA/Hpa II (0.5mg/ml; Biotech; fragment size range from 26-501bp) was loaded into lane 1 as a size standard. The gel was electrophoresed at 80V for approximately 20 minutes. It was then stained for 10 minutes in EtBr (1.5µl 100% EtBr, 30ml water) and visualised under a Hoefer® Mighty Bright™ UV transilluminator.

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

4.9 Fluorescent labelling of markers

Fluorescent detection of alleles using an ABI 373 DNA Sequencer™ with GENESCAN™ allele scoring software was used to accurately size microsatellite alleles. The forward primers for the microsatellites are 22-mer oligonucleotides with either TET (4, 7,2, 7- tetrachloro-6-carboxyfluorescein), HEX (4, 7, 2, 4, 5, 7

– hexachloro-6-carboxyfluorescein) or FAM (6-carboxyfluorescein) molecules chemically bonded to them.

These fluorescently labelled primers are incorporated into the microsatellite markers during amplification with PCR. For loading into a polyacrylamide gel, 1.5µl of PCR sample was mixed with 2.5 µl of formamide, 0.5µl of loading buffer and 0.5µl of TAMRA-labelled internal size standard. The loading buffer and TAMRA-labelled size standard are supplied in the GENESCAN-500 kit™. The size standard is the result of digestion of plasmid DNA with the restriction enzymes *Pst1* and *BstUI*. The resulting DNA fragments are then labelled with TAMRA (N, N, N', N'-tetramethyl-6-carboxyrhodiamine) chemically bonded to it.

4.10 Detection of markers using the ABI 373 DNA sequencer™

A gel mixture of 100ml 40% acrylamide/bisacrylamide at a ratio of 19:1, 420.5g of urea, and 100ml of 10xTBE buffer (890 mM Tris-borate, 2mM NaEDTA.$2H_20$), was made up to 1 litre with distilled water and stored at 4^0C. 30ml of this gel preparation was then mixed with 150µl of 10% ammonium persulphate and 17µl of TEMED. This was dispersed between two glass plates using a 100ml syringe with a 24cm well-to-read distance, and fixed together with bulldog clips. A 50 well square tooth, comb was placed in the top of the gel apparatus, and the gel allowed to set for approximately 2 hours at room temperature. The gel was then pre-run for 5 minutes at 28^0C to optimise the temperature of the gel.

After the pre-run, 3µl of each sample solution was pipetted into the gel wells. The gel was then run for 8-12 hours using filter B. The use of filter B

results in the fluorescent markers displaying the following colours: blue (FAM), green (TET), yellow (Hex) and red (TAMRA). The resulting gel pictures were assembled and analysed with the GENSCAN™ software program. The GENOTYPER™ software program was then used to analyse the data extracted by GENESCAN™, and to assign peaks to microsatellite alleles. The results are presented as a series of peaks with each microsatellite marker resulting in either one peak for homozygotes or two peaks for heterozygotes. The resulting base lengths were then recorded using Microsoft Excel™ version 7.

4.11 Statistical analysis of microsatellite data from the random sample populations

Basic statistical computations, such as the calculation of allele frequencies and of observed and expected heterozygosity, were performed using the GENEPOP program (Rousset 1995; also see Appendix A). The GENEPOP program calculates a range of statistical tests and computations for population genetics research. This program was utilised to calculate HWE probability tests, tests for linkage disequilibrium, the calculation of population correlation coefficients, and the *F* statistics described by Weir and Cockerham (1984). For each of the populations, observed and expected heterozygosities were tested for statistical significance using the χ^2 test.

Analysis of Molecular Variance (AMOVA) was then performed using the ARLEQUIN software program (Excoffier *et al.* 1992). ARLEQUIN is a multi-faceted population genetics software program, freely available on the Internet (see appendix A). ARLEQUIN can process a wide range of genetic information including sequence data, RFLP data and standard frequency data. It can also be

used to perform specific procedures in the treatment of microsatellite data, such as differences in allele size using AMOVA.

Autosomal allele frequencies of the Han and Hui populations were then compared to other populations gathered from the Genome Database (GDB) and Centre d'Études du polymorphisme Humain (CEPH) (see Appendix A). Comparisons were made between the levels of heterozygosity and gene diversity in the study populations and these reference populations. Computation of the Wilcoxon matched pairs signed ranks test was performed using SPSS™ Version 8.0.

Han and Hui Y-chromosome allele frequencies were compared to populations gathered from a database located in the Forensic Laboratory for DNA Research, Department of Human Genetics, University of Leiden. These data are an updated version of information presented in Kayser *et al.* (1997) and de Knijff *et al.* (1997). Comparisons were also made to population data presented in an unpublished article made available by Péréz Lezuan *et al.* Both sets of comparisons were made on the basis of genetic distance calculations.

The calculation of genetic distances was accomplished using the MICROSAT software package (Minch 1998; also see Appendix A). MICROSAT is a program specifically designed for the processing of microsatellite data. MICROSAT generates eight different genetic distances including D_S (Nei 1973) and $\delta\mu^2$ (Goldstein *et al.* 1995). From these distances, unrooted neighbour joining trees (Saitou and Nei 1987) were generated by the PHYLIP version 3.5c software package (Felsenstein 1989). This program was also used to test tree robustness based on a statistical comparison of one thousand bootstrap iterations (Felsenstein 1985).

4.12 Statistical analysis of microsatellite data from the pedigree samples

Pedigree information was collected in China from two families and the pedigree constructed in the Cyrillic™ Version 2.1 software. This software enabled the definition of chromosomal genotypes, and the calculation of pedigree inbreeding coefficients.

Independent genotypes were isolated by analysis of the final constructed pedigrees. Independent individuals were defined as those individuals in the pedigree who represented the genetic founders of the pedigree.

As in the population study, observed and expected heterozygosities were statistically tested using the χ^2 goodness of fit test. Data gathered from independent genotypes and full pedigree data were processed in GENEPOP, in a similar manner to the random sample population data for the calculation of expected heterozygosities, linkage disequilibrium tests and so forth.

Chapter 5

Genetic analysis of the Han and the Hui ethnic groups

5.1 Introduction

The study has been separated into two sections comprising population-based analysis and pedigree-based analysis. In the first section, random sample populations of the Hui and the Han will be compared in terms of within- and between-population genetic variation. The major aim of this part of the study is to gauge the degree to which the genetic structure of the two sample populations matches historical narrative. Therefore emphasis is placed on comparisons of allele distribution patterns, the effects of reproductive isolation, and migration based on the autosomal and Y-chromosome gene pools of both populations.

5.2 Analysis of autosomal frequency distributions

The first step in the study was to compile allele frequency distribution profiles for the ten autosomal loci studied in each of study populations. The allele frequency distributions were similar in both populations, presenting a variety of forms including unimodal, bimodal and multimodal distributions according to the locus studied (see Appendix B). The two populations could however be differentiated by comparing the most frequent allele (MFA) at each loci. Different MFAs were observed in the two populations at a majority of loci, with the Hui generally exhibiting smaller MFAs. The only exception to this pattern was D13S192 (Table 5.1).

The observed patterns of population differentiation were subjected to statistical assessment using an unbiased probability test, according to the method of Raymond and Rousset (1995). The null hypothesis of identical allelic distribution across populations was rejected for eight of the ten autosomal loci (see Table 4.1). Therefore, differentiation between the two populations can be observed by sampling autosomal microsatellite markers.

Further in depth analysis of the allele distributions at each locus exhibited a high level of polymorphism, with the number of alleles per locus ranging from 6 alleles (locus D13S126 in the Hui) to 16 alleles (locus D13S133 in the Han). There was an average of 11.3 alleles per locus recorded for the Han population and 10.1 for the Hui population (Table 5.1). The pattern of allele size distribution also varied slightly between the two populations. The average allele size was calculated as the variance of the number of dinucleotide repeats per locus averaged over each population. An average size variance of 15.52 was recorded for the Han sample population and 16.23 for the Hui (Table 5.1). The statistical significance of these values was assessed by performing a Wilcoxon signed ranks test, which is a non-parametric equivalent of the paired t-test. This test was chosen in preference to the paired t-test as previous studies have demonstrated that the distributions of allele size and frequency show non-normal distributions (Morell *et al.* 1995), and so normalised statistical tests could give unreliable results. Neither the difference in size variance ($p = 0.456$) or number of alleles ($p = 0.062$) showed a significant difference between the Han and Hui. The marginal result recorded for the average allele size variance suggests that the analysis of more loci may have produced a significant result at $p<0.05$.

Even in the absence of statistical significance, the results do concur with available historical evidence on the population structure of the Hui. The paradox of a wide allele size variance but a low number of alleles suggests that the Hui gene pool could have been sourced from a number of diverse origins; but recent endogamy may have accelerated genetic drift thereby reducing the total number of alleles.

Table 5.1 Summary of the distribution of autosomal microsatellite alleles in the Han and Hui random sample populations

		Allele Distribution Data						Exact test
Marker	Population	Allele range (bp)	No. of alleles	MFA*	ASV**	Ho#	He##	p value
D13S126	Han	102-106	7	106	5.81	0.667	0.702	0.0001
	Hui	100-110	6	102	3.50	0.396	0.735	
D13S133	Han	128-189	16	132	31.12	0.621	0.838	0.0001
	Hui	126-183	12	132	20.26	0.490	0.697	
D13S192	Han	87-121	14	97	21.14	0.790	0.823	0.0001
	Hui	93-119	14	103	17.50	0.520	0.883	
D13S270	Han	75-97	10	81	6.00	0.443	0.535	0.0001
	Hui	77-97	8	79	4.67	0.509	0.820	
D15S11	Han	242-266	13	244	15.17	0.618	0.623	0.058
	Hui	240-272	10	244	30.49	0.396	0.478	
D15S97	Han	172-196	10	182	7.12	0.652	0.801	0.005
	Hui	172-188	8	180	13.07	0.373	0.829	
D15S98	Han	145-171	9	157	15.23	0.817	0.779	0.2694
	Hui	131-175	14	153	36.68	0.528	0.808	
D15S101	Han	99-117	12	109	14.08	0.781	0.811	0.0001
	Hui	95-123	10	105	11.57	0.415	0.845	
D15S108	Han	131-161	12	145	29.24	0.556	0.536	0.0276
	Hui	139-165	10	145	12.09	0.490	0.652	
GABRB3	Han	181-201	10	185	10.27	0.430	0.663	0.0001
	Hui	181-201	9	183	12.50	0.320	0.737	
Average	Han	--	11.3	--	15.52	0.637	0.711	--
	Hui	--	10.1	--	16.23	0.457	0.746	--

* MFA = Most frequent allele
** ASV = Allele size variance (variance of allele repeat size)
#Ho = Observed heterozygosity
##He = Expected heterozygosity

5.3 Heterozygosity and gene diversity

The next step in the analysis of the autosomal allele distributions was the calculation and comparison of observed and expected heterozygosity. Observed heterozygosity is defined as the ratio of heterozygotes in a sample population compared to the sum of heterozygotes and homozygotes. Expected heterozygosity is defined as the level of heterozygosity in a population when the population is in Hardy-Weinberg equilibrium. Given p as the frequency of the ith allele at a locus, expected heterozygosity is calculated by the formula:

$$H = 1 - \Sigma p_i^2$$

Equation 5.1 Expected heterozygosity

On average, the Hui exhibited a higher level of expected heterozygosity. 0.746, than the Han 0.711 (Table 5.1, Figure 5.1). However, the difference was not statistically significant according the Wilcoxon signed ranks test ($p = 0.155$). When considered in combination with the higher allele size variance and lower average number of alleles observed in the Hui, the higher expected heterozygosity supports the correlation between the diverse historical origins of the Hui and the genetic structure of the population.

In contrast, the observed heterozygosity in the Hui population was low, 0.456, compared to the Han, 0.657. The difference between these values was significant ($p = 0.03$) and suggests that the Hui are more endogamous than the Han.

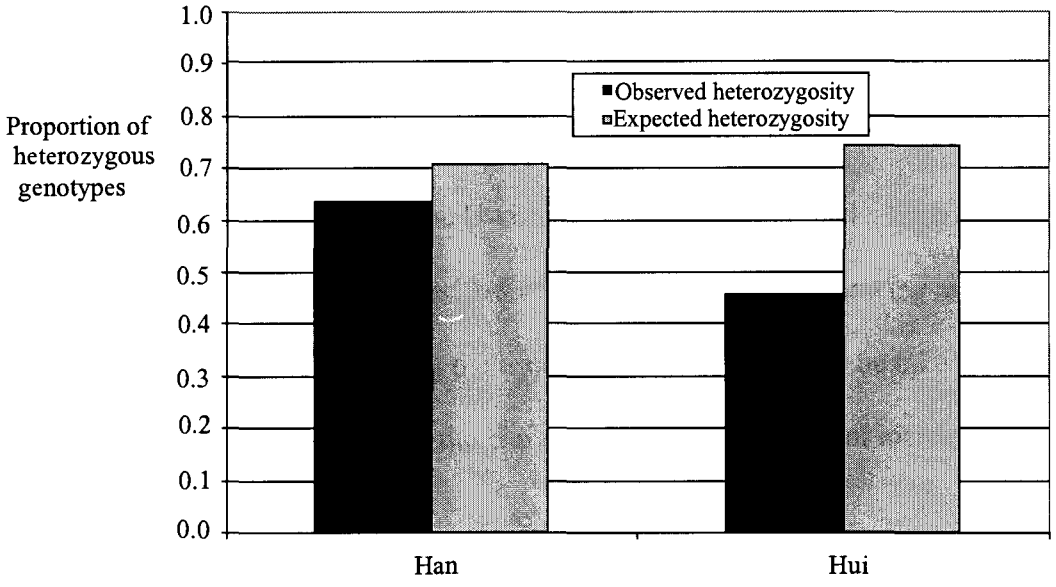
To confirm the differences between the levels of expected and observed hetrozygosity within the Han and Hui, Pearson's χ^2 goodness of fit test was

applied to the data. The results showed that the level of observed heterozygosity in the Hui was significantly smaller ($p<0.001$, d.f. = 9) than the level of expected heterozygosity found in the same population (see Figure 5.1). A significant, but smaller difference was also recorded for the Han ($p = 0.04$, d.f. = 9)

According to population genetics theory, the significant result for both populations may represent either an artefact of unrecognised population subdivision in the population under study and/or the effect of endogamy. As discussed in chapter 2, given the nature of the Han *minzu*, it seems more probable that in the Han population, population subdivision is the principal causative agent.

The lower levels of observed heterozygosity in the Hui may also be due to population subdivision reflected by the diverse origins of the population. However, the difference between gene diversity and heterozygosity levels in the Hui is so large that population subdivision is probably not the sole reason. Another possible explanation is that, in accordance with the available historical evidence, the low level of observed heterozygosity might have resulted from the practice of non-random mating, due to cultural and religious isolation and via preferential consanguinity.

Figure 5.1 Observed and expected heterozygosity in the Han and Hui sample populations



5.4 Hardy-Weinberg equilibrium analysis

To further examine the difference between the levels of expected and observed heterozygosity in both study populations, deviation from HWE expectations was determined along with the direction (heterozygote excess or deficiency) of any observed deviation. The exact probability test, developed by Guo and Thompson (1992) from Fisher's exact test, was used to assess the significance of deviation from HWE. The exact probability test utilises the Monte Carlo Markov chain method to reduce the computational complexity of the exact test permutations, and to provide a confidence interval for testing statistical significance. The U-test developed by Rousset and Raymond (1995), which is an extension of the exact probability test, was used to determine whether there was heterozygote excess or a heterozygote deficiency.

The exact probability test results indicated that, averaged across all loci, both the Han and Hui populations deviated significantly from Hardy-Weinberg expectations (Tables 4.2a and 4.2b). This conclusion is predictable as empirical evidence, such as finite population size and non-random mating, had previously indicated deviations from Hardy-Weinberg equilibrium.

Computation of the U-test showed that the Hui had a high level of heterozygote deficiency, with 9 of the 10 loci surveyed showing significant deviations (D15S11 being the exception). The result may be taken as indicative of non-random mating in the Hui population resulting in an increase in homozygosity.

The results of the exact and U-tests in the Han were not as pronounced as those for the Hui. The exact test showed that only 5 of 10 loci surveyed in the Han exhibited significant deviation from Hardy-Weinberg equilibrium, and the U-test indicated a similar result (Table 5.2a). As with the previous results in the study, the most probable causative explanation would appear to be population subdivision.

Table 5.2a Evaluation of Hardy-Weinberg equilibrium in the Han sample population

Locus	Probability test		U-test	
	p value	S.E.	p value	S.E.
D13S126	0.4881	0.0252	0.0445	0.0120
D13S133	0.0000	0.0000	0.0000	0.0000
D13S192	0.0302	0.0165	0.3295	0.0466
D13S270	0.0033	0.0029	0.0100	0.0047
D15S11	0.4847	0.0557	0.1925	0.0469
D15S97	0.0008	0.0008	0.0022	0.0014
D15S98	0.1860	0.0220	0.2948	0.0341
D15S101	0.0667	0.0277	0.3067	0.0316
D15S108	0.1707	0.0351	0.4742	0.0452
GABRB3	0.0000	0.0000	0.0032	0.0027

Table 5.2b Evaluation of Hardy-Weinberg equilibrium in the Hui sample population

Locus	Probability test		U-test	
	p value	S.E.	p value	S.E.
D13S126	0.0000	0.0000	0.0000	0.0000
D13S133	0.0013	0.0013	0.0000	0.0000
D13S192	0.0000	0.0000	0.0000	0.0000
D13S270	0.0000	0.0000	0.0000	0.0000
D15S11	0.0733	0.0223	0.0889	0.0269
D15S97	0.0000	0.0000	0.0000	0.0000
D15S98	0.0000	0.0000	0.0000	0.0000
D15S101	0.0000	0.0000	0.0000	0.0000
D15S108	0.0000	0.0000	0.0000	0.0000
GABRB3	0.0000	0.0000	0.0000	0.0000

5.5 Gametic association

As the results to date have appeared to indicate significant non-random mating patterns in both study populations due to population stratification, the Han and Hui sample populations were subject to testing for gametic association. If no population stratification exists in a population, then each separate locus would segregate independently regardless of the effect of the presence of other loci. In theory, non-random mating would nullify the assumption of independence of loci, and create significant non-random associations between loci. The presence of non-random associations is termed linkage disequilibrium. However, as discussed in Chapter 3, the use of the word linkage may not be accurate as the presence of linkage disequilibrium can occur between alleles that are not necessarily physically linked. Therefore it is proposed that the term gametic association be used to describe this phenomenon, as the major interest of the study lies in assessing the effect of non-random mating on allele frequencies, rather than disease association and gene mapping, where the term linkage has a more direct application.

To investigate if gametic associations are present in the autosomal data from the Han and the Hui, the exact probability test was used, as in the previous tests for population differentiation. Exact tests for gametic association depicted significant p values for three pairs of loci in the Han population and two pairs of loci for the Hui, from a total of 45 comparisons. The proportion of significant results expected by chance alone is approximately 2 (expected from type I error at $\alpha = 0.05$). Therefore, in the ten loci surveyed the Han exhibited a slightly elevated number of deviations from genotypic equilibrium while the Hui did not.

The Han result provides additional evidence in support of the hypothesis of population subdivision present within the Han *minzu*. The three locus pairs at which significant gametic association were observed were D13S192/D13S270 ($p = 0.019$), D13S192/GABRB3 ($p = 0.008$) and D13S270/D15S98 ($p = <0.001$). It appears highly improbable that there is linkage in any of these cases, as the first pair are 38.691 cM apart (Figure 4.1) and the other two pairs are on different chromosomes. It is however of interest that one of the two significant gametic associations in the Hui was also D13S192/GABRB3 ($p = 0.021$).

Pritchard and Rosenberg (1999) suggested that the presence of a significant number of unlinked loci pairs was indicative of the presence of population subdivision therefore it could be concluded that the Han of Liaoning are composed of several or even multiple subpopulations. However the number of loci surveyed are small and they are restricted to just two chromosomes, and a genome-wide screen employing many more loci would be needed to produce a more definitive result.

One possible explanation for the non-significant result obtained with the Hui is the historical admixture of Han females. In combination with the diverse male founding populations, this may have resulted in multiple recombination events through the generations. Equally, as in the Han, the result may principally reflect the small number of loci tested. Pritchard and Rosenberg (1999) recommended the study of at least 15 unlinked microsatellites to test for population stratification using a gametic association approach. Only ten microsatellites were surveyed in the present study.

Table 5.3a Gametic association in the Han

Locus	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108
D13S133	0.669								
D13S192	0.139	0.988							
D13S270	0.104	0.905	[illegible]						
D15S11	0.774	0.742	0.578	0.889					
D15S97	0.061	0.952	0.359	0.321	0.688				
D15S98	0.138	0.881	0.708	[illegible]	0.721	0.623			
D15S101	0.518	0.242	0.712	0.454	0.748	0.820	0.971		
D15S108	0.432	0.994	0.921	0.098	0.846	0.981	0.146	0.199	
GABRB3	0.571	0.604	[illegible]	0.750	0.162	0.332	0.267	0.156	0.369

Table 5.3b Gametic association in the Hui

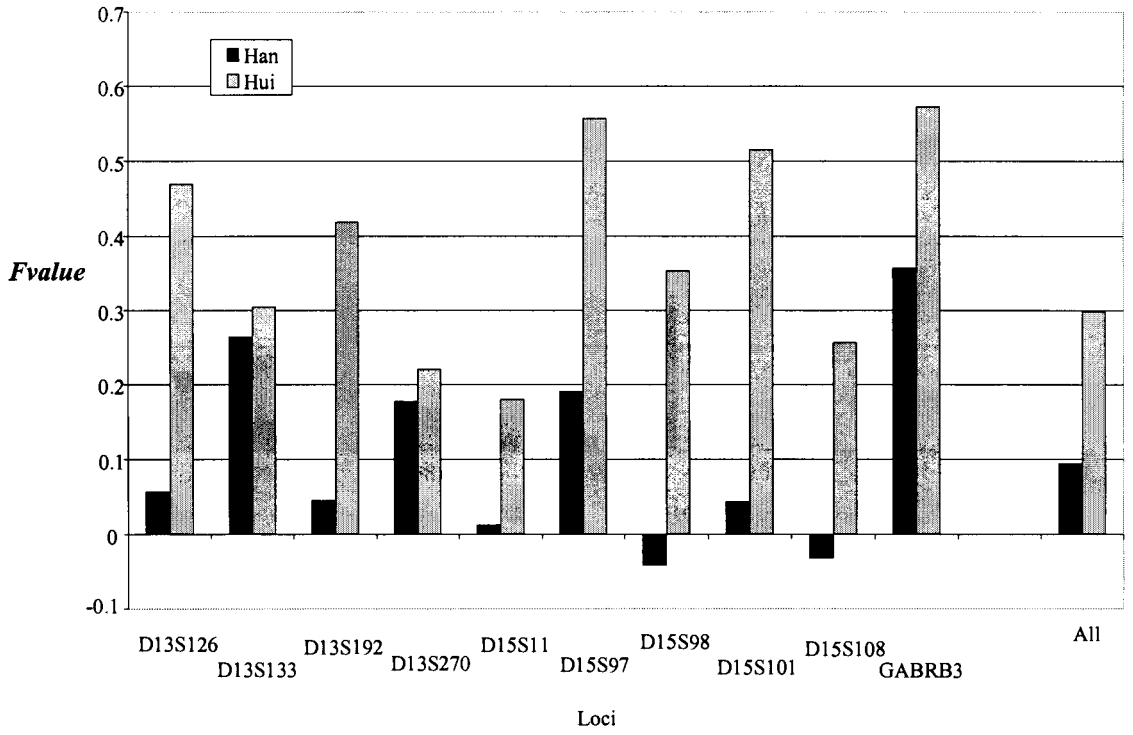
Locus	D13S126	D13S133	D13S192	D13S270	D15S11	D15S97	D15S98	D15S101	D15S108
D13S133	0.202								
D13S192	0.399	0.084							
D13S270	0.764	0.353	0.971						
D15S11	0.100	0.347	0.317	0.919					
D15S97	0.747	0.707	[illegible]	0.804	0.384				
D15S98	0.333	0.462	0.768	0.707	0.873	0.088			
D15S101	0.368	0.619	1.000	0.383	0.773	0.895	0.769		
D15S108	0.264	0.541	0.649	0.775	0.163	0.638	0.995	0.919	
GABRB3	0.070	0.650	[illegible]	0.926	0.270	0.624	0.664	0.257	0.960

5.6 Allelic correlation coefficient

Since the assessment of gametic association was unsuccessful in quantifying non-random mating in the study populations, an alternative method was utilised to assess the within-population genetic structure of the Han and the Hui. The method chosen was estimation of the allelic correlation coefficient (f), which measures the correlation of genes within individuals within populations (see section 3.3.6). This coefficient was derived according to the partitioning of analysis method devised by Weir and Cockerham (1984) and was tested using the GENEPOP software. In effect, the parameter is a multiallelic version of the f parameter described in equation 3.6. As f is a parameter, rather than a statistic, it is more statistically robust than Wright's original F_{IS} (Weir and Cockerham 1984).

High between-locus levels of variation were observed in the correlation coefficients calculated for the Han population. The f estimates varied between -0.041 and 0.357 with an average value of 0.100. In contrast, f values calculated for the Hui population, ranged from 0.181 to 0.557 with a mean of 0.400. The results indicate that the Hui have a greater number of alleles that are identical by state than the Han population, a finding in keeping with the isolated nature of the population. The variance of f coefficients between negative and positive values in the Han indicates a more complex situation, providing further presumptive evidence for the presence of population substructure.

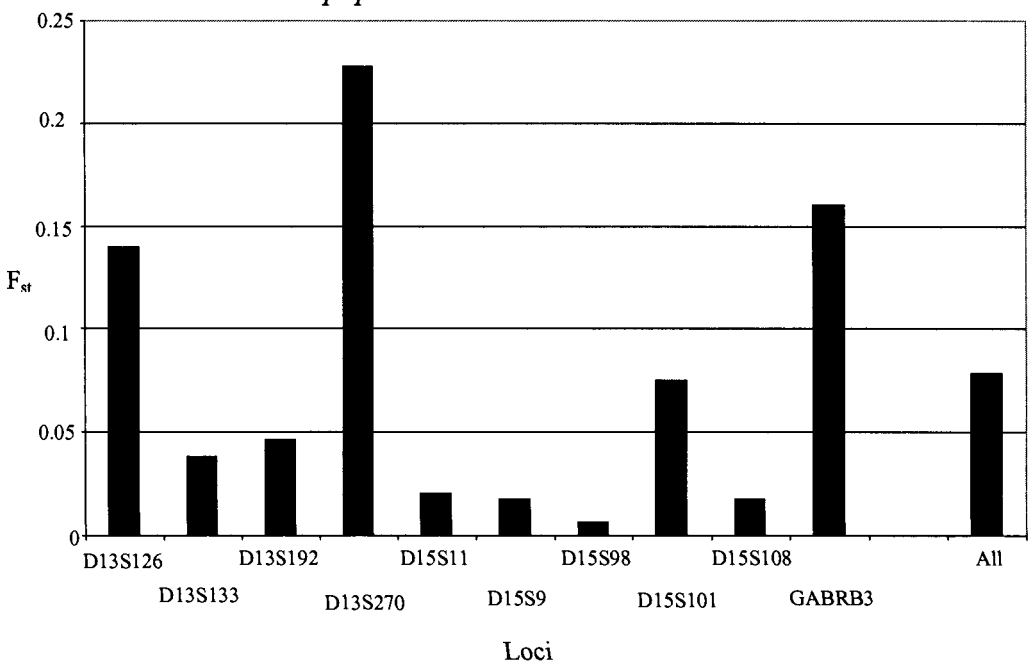
Figure 5.2 Correlation coefficients of the Hui and Han sample populations



5.7 F_{ST}

The next step in the study was to establish the degree of genetic variance between the two populations. This was attempted by calculation of parameters analogous to Wright's F_{ST} according to the method of Weir and Cockerham (1984), once again using the GENEPOP software. The average F_{ST} calculated for the 10 loci surveyed was 0.0793. However, individual F_{ST} values exhibited large variance from locus to locus with values ranging from 0.0071 to 0.2283. For greater statistical reliability of the average F_{ST} value the analysis of additional loci would be helpful.

Figure 5.3 F_{ST} calculated from Hui and Han sample populations



5.8 AMOVA calculated for the autosomal loci

The analysis of molecular variance (AMOVA) was chosen to assess between-population genetic variation. The hierarchical nature of AMOVA allows a statistically robust assessment of the variance seen in individual chromosomes within and between populations. For this study, data from chromosomes 13 and 15 were tested individually and as a pooled data set.

In all tests the between-population variance is represented as Φ_{ST}. The significance of the Φ_{ST} statistic was assessed by the method described by Excoffier *et al.*(1992), where the original calculated Φ_{ST} value is compared to a distribution of Φ_{ST} values generated from 10,000 random permutations of the original data set. If the calculated Φ_{ST} value is larger than 95% or more of the generated Φ_{ST} values, then it is deemed significant.

As mentioned in section 3.3.4, two methodologies pertain when conducting AMOVA on microsatellite data, the difference in the number of alleles and the sum of squared allele size difference. According to Mickalakis and Excoffier (1996) the first method is analogous to F_{ST} analysis (Weir and Cockerham 1984) which earlier was found to be of limited application (Section 5.7). Instead, the sum of squared allele size difference method was utilised. Unlike F_{ST} analysis, it is not subject to sample size bias, and as such, is more statistically robust (Goldstein *et al.* 1995).

AMOVA analysis showed no significant between-population variance for the separate chromosome data (Table 5.4), possibly due to the small number of loci surveyed. By comparison, the data from all ten loci produced a significant between-population result ($p<0.001$). The proportion of variance, equivalent to an

F_{ST} value of 0.0463, was much smaller than the global average F_{ST} of 0.15 reported by Babujani *et al.* (1997), and may reflect the admixture of Han females within the Hui population.

Table 5.4 Analysis of the molecular variance (AMOVA) of autosomal markers in the Han and Hui random sample populations

Chromosome	Φ_{ST} (%)	Φ_{IS} (%)	Significance of Φ_{ST}	
			P	S.E (+/-)
13	2.09	97.91	0.06	0.008
15	1.14	98.86	0.06	0.008
13 + 15	4.63	95.37	<0.001	< 0.001

5.9 Reference population comparisons

In order to gain some perspective on the population structure of the Han and Hui, comparisons were made with a Caucasian reference population composed of CEPH (D13S126, D13S270, D15S11-GABRB3) and GDB data (D13S133 and D13S192) Given their Western European origins, it assumed that neither data sets are from endogamous communities. As with the comparison between the Han and the Hui, these tests were conducted using the Wilcoxon signed ranks test. A summary of the reference data is shown in Table 5.5.

The reference population produced results similar to the Han population. The mean number of alleles was the same at 11.3, and the mean ASV values were very similar ($p = 0.420$) with 15.495 in the reference population and 15.518 in the Han random sample population. It can be concluded from these comparisons that the Han exhibit a similar level of genetic diversity to that of this composite

Caucasian population. Therefore, a possible interpretation is that the Han population in Liaoning represents a similarly broad grouping.

The values for the Hui were lower than for the reference population (see Table 5.1). The ASV values for the reference data and the Hui were not significantly different ($p = 0.401$). The comparisons of mean number of alleles also were non-significant, but the result was marginal ($p=0.054$) again suggesting some degree of population endogamy.

The observed and expected heterozygosity levels in the reference population were very similar, 0.745 and 0.747 respectively. Comparisons of the reference population with Han observed and expected heterozygosity levels showed that expected hetrozygosity levels were not significantly different ($p = 0.114$) but that observed heterozygosity levels in the Han were significantly lower than levels observed in the reference population (one tailed $p = 0.021$). Similar conclusions were drawn from comparisons of expected and observed hetrozygosity levels between the Hui and the reference population. The expected heterozygosity levels were not significantly different ($p = 0.721$) but the observed heterozygosity levels in the Hui were significantly smaller than those found in the reference population (one tailed $p = 0.001$).

In conclusion, both the Han and the Hui exhibited significantly lower observed heterozygosity levels than the reference population. This effect can most convincingly be ascribed to population subdivision in both of the Chinese populations.

Table 5.5 Summary of allele distributions of autosomal microsatellite markers in the reference population

Marker	Allele range (bp)	No. of alleles	MFA*	ASV**	#Ho	##He
D13S126	100-112	7	110	4.66	0.650	0.677
D13S133	132-187	15	136	21.97	0.800	0.820
D13S192	89-123	15	103	24.92	0.880	0.900
D13S270	75-95	6	89	6.17	0.679	0.680
D15S11	238-262	11	238	12.96	0.614	0.643
D15S97	168-186	10	248	9.17	0.899	0.811
D15S98	141-175	17	155	26.56	0.852	0.860
D15S101	101-133	9	123	21.14	0.763	0.781
D15S108	141-161	10	157	12.22	0.570	0.563
GABRB3	171-201	13	185	15.18	0.743	0.734
Average	--	11.3	--	15.495	0.745	0.747

* MFA = Most frequent allele
** ASV = Allele size variance (variance of allele repeat size)
#Ho = Observed heterozygosity
##He = Expected heterozygosity

5.10 Analysis of Y-chromosome allele frequency distributions

Y-chromosome diversity in the two study populations also was analysed, with the patterns of Y-chromosome allele variation observed in the Hui additionally assessed to see if they concurred with the hypothesis of male-directed gene flow.

The markers DYS19, DYS388, and DYS390 showed similar distributions in the Han and the Hui, with both populations sharing the same MFA (Table 5.7). By comparison, at the four other markers examined, DYS389I, DYS389II, DYS392 and DYS393, variant MFAs were observed in each population. As with the autosomal data, the level of population differentiation exhibited at each locus was statistically assessed using the exact test. The results showed statistically significant differentiation at four of the seven Y-chromosome loci (Table 5.7).

Further analysis of the Y-chromosome allele distributions indicated that the Hui had a lower average number of alleles per locus, 4.3, and a lower mean allele size variance, 2.2, than the Han, 5.3 and 3.1 respectively. The Wilcoxon signed ranks test indicated that the variance of allele size difference in the Hui was significantly lower than the corresponding Han value (one tailed $p = 0.030$). A similar difference did not, however, exist between the number of alleles in the two populations (one tailed $p = 0.065$).

Table 5.6 Summary of allele distributions of Y-chromosome markers in the Han and Hui

		Allele distribution data					Exact test
Marker	Population	Allele range (bp)	No. of alleles	MFA*	ASV**	Gene diversity	p value
DYS19	Han	186-202	5	194	3.5	0.663	0.905
	Hui	190-202	4	194	1.7	0.676	
DYS388	Han	126-144	6	129	4.7	0.649	0.001
	Hui	129-144	6	129	3.5	0.649	
DYS389I	Han	245-261	5	249	2.5	0.590	0.110
	Hui	249-257	3	253	1.0	0.612	
DYS389II	Han	361-381	6	365	3.5	0.747	0.032
	Hui	365-377	4	369,373	1.7	0.727	
DYS390	Han	203-223	6	215	3.5	0.768	0.002
	Hui	203-223	6	215	3.5	0.716	
DYS392	Han	251-269	5	257	2.5	0.694	0.014
	Hui	251-260	3	251	2.3	0.547	
DYS393	Han	120-132	4	120	1.7	0.592	0.067
	Hui	120-132	4	124	1.7	0.665	
Average	Han	--	5.3	--	3.1	0.672	--
	Hui	--	4.3	--	2.2	0.656	--

* MFA = Most Frequent Allele
** ASV = Allele size variance (variance of allele repeat size)

5.11 Y-chromosome gene diversity

To accurately compare Y-chromosome and autosomal genetic diversity in the Han and Hui, the gene diversity of both populations was calculated. Gene diversity is mathematically akin to expected heterozygosity but, as the Y-chromosome is effectively haploid, this value represents the probability that two randomly chosen alleles from the same population are different.

The gene diversity of Y-chromosome markers in the Han and Hui were 0.672 and 0.656 respectively (Table 5.7). While gene diversity in the Han was slightly higher than in the Hui, the difference was not statistically different, according to the Wilcoxon signed ranks test ($p = 0.452$). When compared to average gene diversities derived for different European populations, such as the Basques, 0.435 (Péréz Lezuan *et al.* 1997b), a German population, 0.389, and a Dutch population, 0.435, (Roewer *et al* 1996), the Hui and Han Y-chromosome diversities were very high. This suggests that the Han and Hui contain a more diverse range of Y-chromosome haplotypes then these European populations.

It was proposed by Péréz-Lezuan *et al.* (1997b) that it was possible to directly compare Y-chromosome and autosomal gene diversity by transforming Y-chromosome gene diversity using the following formula:

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

Equation 5.2 Ratio of autosomal to Y-chromosome gene diversity

where D_{au} is autosomal gene diversity and D_y is Y-chromosome diversity. This method assumes that the effective number of Y-chromosomes is one quarter the number of autosomal chromosomes. From equation 5.1 an equivalent autosomal

gene diversity of 0.886 was calculated for the Hui, and 0.895 for the Han, from their respective Y-chromosome gene diversities. These values are higher than the actual expected heterozygosity values of 0.754 and 0.713 respectively. By comparison, the same analysis performed on a Basque population resulted in the adjusted Y-chromosome gene diversity being in close agreement with autosomal gene diversity (Péréz-Lezuan *et al.* 1997b).

The most parsimonious explanation for the contrast between the adjusted Y-chromosome diversity and the autosomal gene diversity in the Han and the Hui is the presence of population substructure. The Basque population for which equation 5.1 was created is an isolated endogamous population with no history of population substructure (Péréz-Lezuan *et al.* 1997b). However, the results obtained from the analysis of autosomal diversity in the Han and the Hui clearly exhibited evidence of population substructure (Sections 5.1 – 5.9). Therefore the simple model used by Péréz-Lezuan *et al.* (1997b) is not appropriate for comparative studies in the Han and the Hui.

5.12 Y chromosome haplotype analysis

The advantage of Y-chromosome studies is that direct haplotype analysis can be readily performed, and so the next stage of the study was to undertake this step. From the Han, 75 complete haplotypes consisting of all markers (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS392, and DYS393) could be extracted from a total of 102 samples. As noted in section 4.7, in some cases specific markers could not be amplified from all Han samples, possibly due to partial sample degradation. From the 26 male Hui samples, 18 complete haplotypes could be defined, two of which proved to be identical.

A comparison of the Han and the Hui haplotypes showed no shared haplotypes between the two populations, demonstrating complete differentiation of the two populations. The power of Y-chromosome haplotype analysis is due to the fact that a majority of the chromosome does not undergo recombination. This allows researchers to create haplotype networks which allows male gene flow to be traced (Jobling 1995). The construction of networks using microsatellite haplotypes, however, presents computational problems. A study by Roewer *et al.* (1996) showed that in two closely related European populations, German and Dutch, there were $3x10^{27}$ equally parsimonious haplotype networks based on just four Y-chromosome microsatellite markers (DYS19, DYS389I, DYS389II and DYS390).

A similar analysis on more heterogenous populations, such as the Han and the Hui, using a larger number of markers would logically result in an even greater number of possible networks. Therefore, in the present study a more feasible method would be to perform a broader analysis of haplotype diversity in the Han and the Hui using AMOVA analysis.

In summary, AMOVA analysis of Han and Hui Y-chromosome haplotypes showed that 10.31% of total molecular variance could be ascribed to between-population variance (Φ_{ST}= 0.1031, p = 0.001), which is more than double the size of the corresponding autosomal value (Φ_{ST}= 0.0463, p <0.001) (Tables 5.5 and 5.7).

Besides comparing the seven locus haplotypes, a shorter six locus haplotype based on markers DYS19, DYS388, DYS389I, DYS390, DYS392 and DYS393 was also tested by AMOVA. The reason for this step was that the DYS389II fragment comprises the length variation in DYS389I plus three

additional stretches of tetranucleotide repeats (Rolf *et al.* 1998, Péréz Lezuan *et al.* 1999). Therefore, in the seven locus haplotype, DYS389I is effectively counted twice which could suggest an apparently more homogenous haplotype pool than actually is present in the two populations.

As found with the seven locus haplotypes, no haplotypes were shared between the Han and Hui, thus confirming the complete differentiation between the populations. AMOVA for the six loci haplotype demonstrated a significant between-population molecular variance amounting to 13.99% of total molecular variation (Φ_{ST} = 0.1399, p <0.001), an increase of 4% from the seven locus haplotype and demonstrating the homogenising effect of counting both DYS389I and DYS389II in a haplotype.

Table 5.7 Analysis of molecular variance (AMOVA) of seven and six locus Y chromosome haplotypes

Haplotype	Φ_{ST} (%)	Φ_{IS} (%)	Significance of Φ_{ST}	
			P	S.E (+/-)
Seven allele	10.13	89.87	0.001	0.001
Six allele	13.99	86.01	<0.001	0.001

The Φ_{ST} value of 0.1399 is again more than double than the autosomal Φ_{ST} value of 0.0463. However, a direct comparison of these two values may not be appropriate as the Y-chromosome result was based on haplotypic variance whereas the autosomal reflects genotypic variance. In addition, the Y-chromosome is effectively haploid while the autosomal chromosomes are diploid. In conclusion, an indirect analysis would be more appropriate.

To this end, the Φ_{ST} value obtained from the study was compared to other Φ_{ST} values reported from comparisons of four different European populations (de

Knijff *et al.* 1997). The European Φ_{ST} values ranged from 0.007 from a comparison of Dutch and Swiss populations to 0.0812 between Dutch and German populations. These values are obviously lower than the Φ_{ST} value of 0.1399 in the present study, indicating that the Han and Hui are comparatively less related to each other than the European populations tested. The findings are thus in keeping with historical evidence showing that the Han and the Hui originated from different male founding populations.

5.13 Comparison of Y-chromosome haplotypes to reference populations

The calculation of genetic distances, and the construction of unrooted neighbour–joining trees provides a further alternative to haplotype network construction for population comparisons based on Y-chromosome markers. This analysis was undertaken by comparing Han and Hui data with information sourced from the database at the Centre for Forensic Genetics, Leiden University and Péréz -Lezuan *et al.* (1999). The comparisons were based on the calculation of *Ds* distance (Equation 3.24), $(\delta\mu^2)$ genetic distance (Equation 3.25), and R_{ST} (Equation 3.22), from which unrooted neighbour-joining trees were constructed.

The first comparisons were with broad level continental data from the Leiden database (Figure 5.4 a-c) with the Han and the Hui matched with populations classified as Southeast Asian, Northeast Asian, Indian and European. All the distances calculated resulted in trees that had one recurring property, the shortest distance was the distance between the Han and the Hui. The small distance between the Han and Hui could be explained on the basis of internal population structure. The reference populations were composite populations, amalgamated from different sample populations collected and assessed by different laboratories and simply grouped into their region of origin. As a result

little or no regard was paid to the anthropology, history or demography of the populations included in each of the population groups. By contrast, classification of the Han and the Hui populations were primarily based on cultural and religious grounds. Therefore comparison of the two different population types may not be meaningful.

With this caveat in mind, two specific populations were isolated from the broad population groups analysed by de Knijff *et al.*(1997) using information contained in an appendix published by the authors. The populations chosen were a population termed Chinese, extracted from the Southeast Asian group, and a population termed Mongolian, extracted from the Northeast Asia group. The same European sample population analysed in the first tree construction (Figure 5.4) was used for the analysis, as it has been shown (Cavalli Sforza et *al.* 1994, de Knijff *et al.* 1997, Péréz Lezuan *et al.* 1997b) that genetic variation between European populations is significantly smaller than in the other major continental population groups. Four Central Asian populations also were compared to the Hui and Han (Figure 5.5a-c) using data sourced from Péréz Lezuan *et al.* (1999). These were the Uigur, Kazakh, and two Kirghiz populations collected in the Republics of Kirghizstan and Kazakhstan on the border with China.

The three consensus trees generated are very similar. In all three trees the Central Asian population are clustered together and separated from the other populations. Therefore, for the purposes of this study, these Central Asian populations can be considered as one. Given this assumption, it can be concluded that the first two trees, D_s and D_{DM} (Figures 5.5a and 5.5b), have the same physical structure, with the D_{DM} tree being more statistically robust. The third tree separates the populations in three clusters, European, Central Asian, and East

Asian. This seems to be the most credible tree construction, based on evidence from historical and anthropological data (Hammer *et al.* 1998).

However, all trees displayed the Han and Hui on the same branch as in the first constructions, with lesser genetic variation identified between them with any of the other populations. The results thus suggests the male Han and Hui of Liaoning may be more closely related than first thought. Alternatively, in the absence of appropriate data from other East and West Asian populations, the true genetic lineages of the Han and the Hui cannot be resolved from the present phylogenetic analyses.

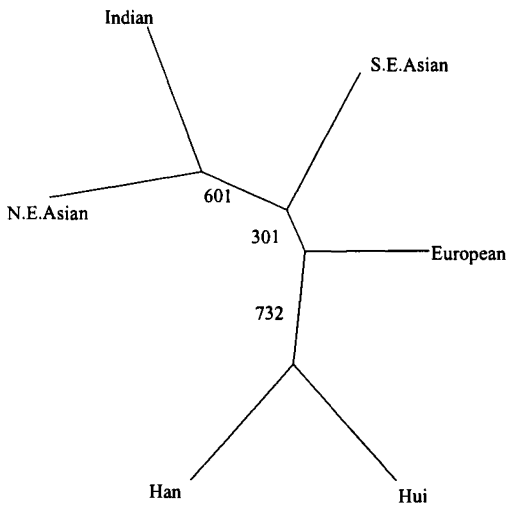
It can also be argued that the small distance between the Han and the Hui may reflect similarly complex population structures rather than shared genetic ancestry. The mean gene diversity for the combined Central Asian populations was 0.432 (Péréz-Lezuan *et al.* 1999), a small value compared to gene diversities of 0.672 and 0.656 for the Han and Hui. As Nei's distance, D_S, is based on differences in gene diversity (Equation 3.24), the trees constructed from this distance would reflect the clustering of populations around similar heterogeneity and not necessarily be based on a coalescent ancestry.

The SMM distances used (D_{DM} and R_{ST}) (Figures 5.5 b, c) are based on allele size difference and, in theory, would more accurately define coalescent ancestry. However, in order to define ancestry, it must be assumed that one population is a founder of another population and the population groups used in the present study are too broad to permit any such assumption. For example, both the Han and the Hui have multiple ancestries, demonstrated by their large gene diversities, while the Central Asian population grouping appears more restricted, as indicated by the low diversity and the multiple instances of shared haplotypes

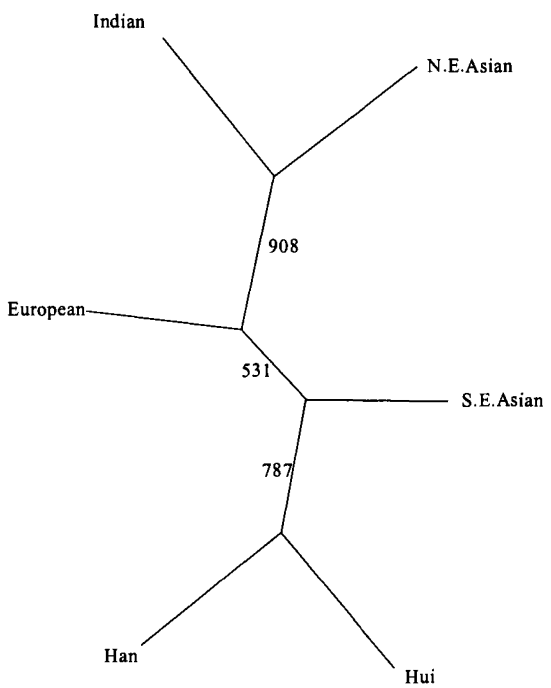
(Péréz-Lezuan *et al.* 1999). Therefore, unless the Han and Hui populations are to be further defined into smaller groups based on shared ancestry, the use of SMM distances in the present study may not be accurate for tracing male gene flow.

Figure 5.4 (a) –(c). Unrooted Neighbour-Joining trees showing phylogenetic relationships between five populations based on five different genetic distances. The numbers at each node indicate the number of trees out of 1000 bootstrap replicates with such a node.

(a) D_S



(b)D_{DM}



(c) R_{ST}

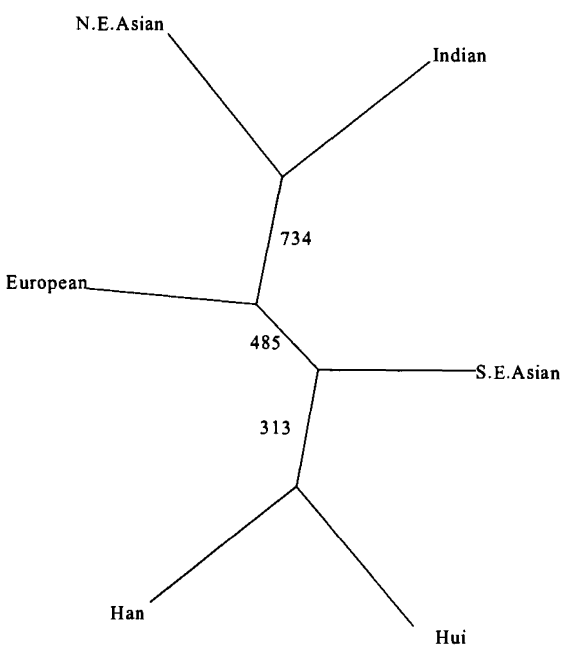
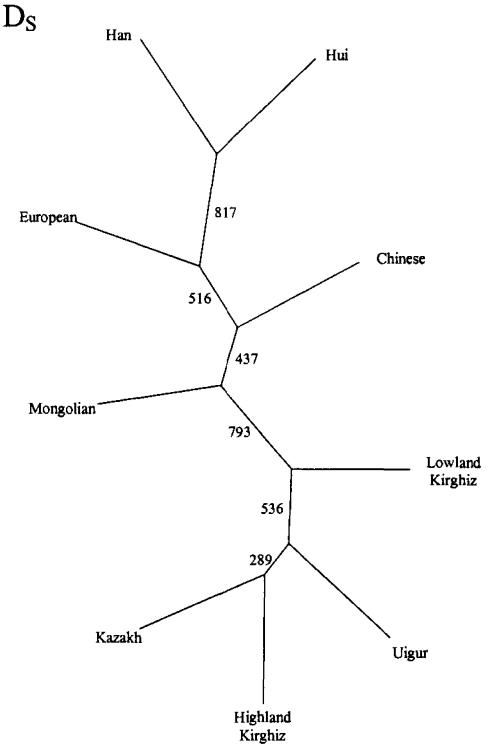
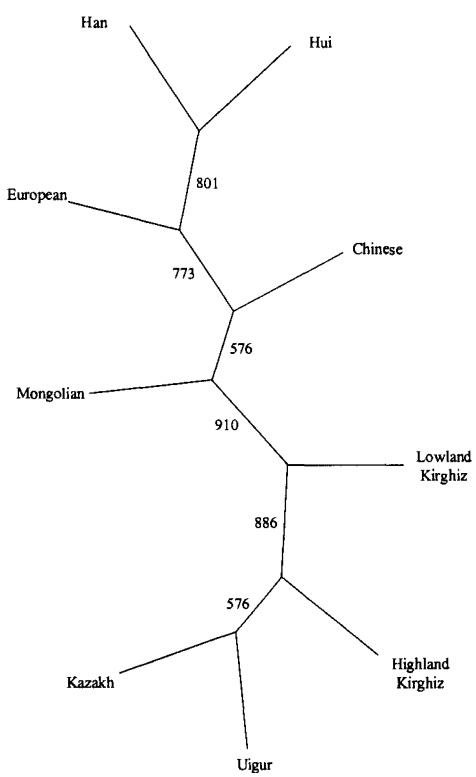


Figure 5.5 (a)-(c) Unrooted Neighbour-Joining trees showing phylogenetic affinities between nine populations using five different genetic distances. The numbers at each node indicate the number of trees out of 1000 bootstrap replicates with such a node

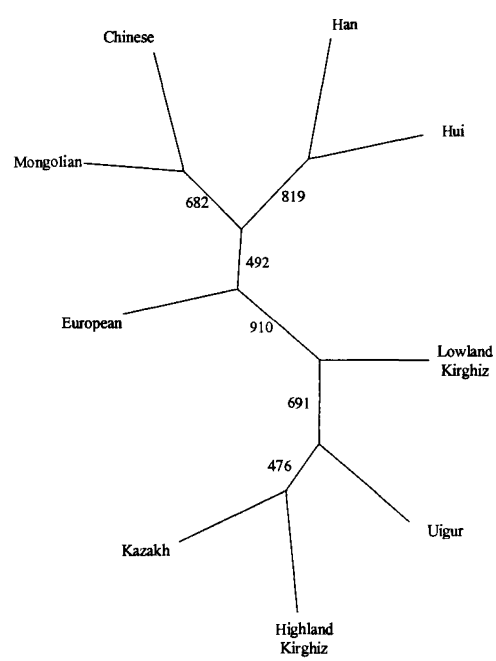
(a) D_S



(b) D_{DM}



(c) R_{ST}



Chapter 6

Analysis of the effects of consanguinity on genetic variation in the Hui of Liaoning Province

6.1 Introduction

Comparisons of inter-population structure showed that the two populations differed genetically, with the Hui revealing a greater deficiency in heterozygosity than the Han, probably ascribableto their greater levels of endogamy. Since the major genetic differences appeared to result from within-population differentiation, the internal structure of the Hui population was further explored via pedigree analysis.

As discussed in Chapter four, two Hui pedigrees were constructed from data gathered in P.R. China and the blood samples were taken from these individuals. This provided an opportunity to study the effects of consanguinity on genetic diversity in the Hui and thus examine the internal structure of the population. The analysis of the pedigree data was conducted by calculation of pedigree inbreeding coefficients, an analysis of allele number and size distribution, calculation of the observed and expected heterozygosity levels, investigation of Y-chromosome marker diversity, and calculation of the allelic correlation coefficients.

6.2 Pedigree structure and identity by descent

The number of individuals sampled from the Wang pedigree numbered 31 of a total of 81 individuals identified, and for the Wu 13 of 17 individuals identified were sampled. The pedigrees were constructed from information supplied by local religious leaders and from collaborating researchers in P.R. China (appendix D). Genotypes for individual M (mother), in the Wang pedigree were inferred from the genotypes of her children and by comparison with her spouse.

It is evident that in both pedigrees there is an appreciable prevalence of consanguineous marriage (Appendix D). However, it was impossible to calculate an inbreeding coefficient from the Wu pedigree, because although consanguineous marriages were identified no information was available on the actual degree of each relationship. Of the 6 consanguineous marriages identified in the Wang pedigree, specific relationships had been defined for three of the marriages; F = 0.0625, 0.0156 and 0.0039, equivalent to a first cousin, second cousin and third cousin marriage respectively. From these figures, a mean inbreeding coefficient (α) of 0.001 was calculated. Clearly this represents an under-estimate as the three other known consanguineous unions could not be included in the calculation and the cumulative inbreeding coefficient from the most recent common ancestor (MRCA) also was unavailable because of a lack of data on the relevant individuals.

Due to this lack of data, an estimated maximum mean pedigree inbreeding coefficient was calculated based on the assumption that the most likely maximum F value for any individual would be 0.0625. This figure was chosen as the maximum possible on the basis that first cousin unions (F = 0.0625) are the most common marriages among consanguineous communities (Bittles and Neel 1994), including those in P.R. China (Bittles 1998). Using this assumption, estimated maximum mean pedigree inbreeding coefficients (α) of 0.003 and 0.007 were derived for the Wang and Wu pedigrees respectively. These results are similar to those obtained by means of household surveys from Hui populations in Gansu, Guandong and Guizhou, PR China, which ranged from 0.0012 to 0.0065 (Wu 1987). The figures are high compared to values derived from similar studies of consanguinity in Han populations, which recorded mean inbreeding coefficients

from 0.0003 to 0.0045 depending on the region of China surveyed (Du *et al.* 1981; Wu 1987, Zhan *et al.* 1992).

6.3 Autosomal allele distributions

Analysis of the allele distributions shows that the pedigrees exhibit a smaller number of alleles per locus and a smaller allele size variance than the random Hui sample population. The Wang and Wu had an average number of alleles of 6.0 and 4.3 respectively and an average size variance of 13.9 and 12.0 respectively (Table 6.1) by comparison with an average of 10.3 alleles per loci and an average size variance of 16.2 in the Hui random sample population (Table 5.1).

Calculation of statistical significance using the Wilcoxon signed ranks test showed no significant difference in allele size variance between either the Wang or Wu pedigrees and the Hui random sample population (one tailed $p = 0.187$ and 0.063 respectively). However, there was a significant difference in the number of alleles between the Wang or the Wu and the Hui random sample ($p = 0.001$ and 0.02 respectively).

As the pedigrees would predicably have a more constricted number of genotypes, this contraction in allele number was not unexpected. Nonetheless the average size variance values calculated for the pedigrees show that, while allele numbers were low, there was still a wide range of allele sizes present at most loci. As in the unrelated sample data, this finding may be indicative of residual genetic diversity due to the diverse genetic origins of the Hui.

Table 6.1 Summary of allele distributions of autosomal microsatellite markers in the Wang and Wu pedigrees

Marker	Population	Allele range (bp)	No. of alleles	MFA*	ASV**	Ho#	He##
D13S126	Wang	104-122	5	106	2.5	0.548	0.527
	Wu	106-122	3	106	2.3	0.462	0.492
D13S133	Wang	133-185	4	133	44.7	0.613	0.621
	Wu	133-177	5	175	15.7	0.846	0.723
D13S192	Wang	95-123	10	107	20.9	0.935	0.861
	Wu	97-115	7	97	10.6	0.917	0.797
D13S270	Wang	81-91	3	81	7.0	0.645	0.529
	Wu	81-91	3	81	7.0	0.154	0.151
D15S11	Wang	242-272	7	244	29.6	0.613	0.544
	Wu	244-260	4	244	11.3	0.769	0.578
D15S97	Wang	176-188	6	172,182	4.7	0.800	0.721
	Wu	170-182	4	176	7.0	0.385	0.443
D15S98	Wang	143-173	7	151	6.7	0.731	1.000
	Wu	145-171	5	153,157	22.8	0.733	0.778
D15S101	Wang	103-113	6	107	3.5	0.903	0.819
	Wu	101-113	6	107	5.4	0.846	0.778
D15S108	Wang	143-159	5	157	9.7	0.733	0.773
	Wu	143-157	3	143	10.3	0.692	0.557
GABRB3	Wang	179-197	7	185	9.6	0.774	0.667
	Wu	179-199	3	183	28.0	0.770	0.540
Average	Wang	--	6	--	13.9	0.729	0.706
	Wu	--	4.3	--	12.0	0.657	0.584

* MFA = Most frequent allele
** ASV = Allele size variance (variance of allele repeat size)
#Ho = Observed heterozygosity
##He = Expected heterozygosity

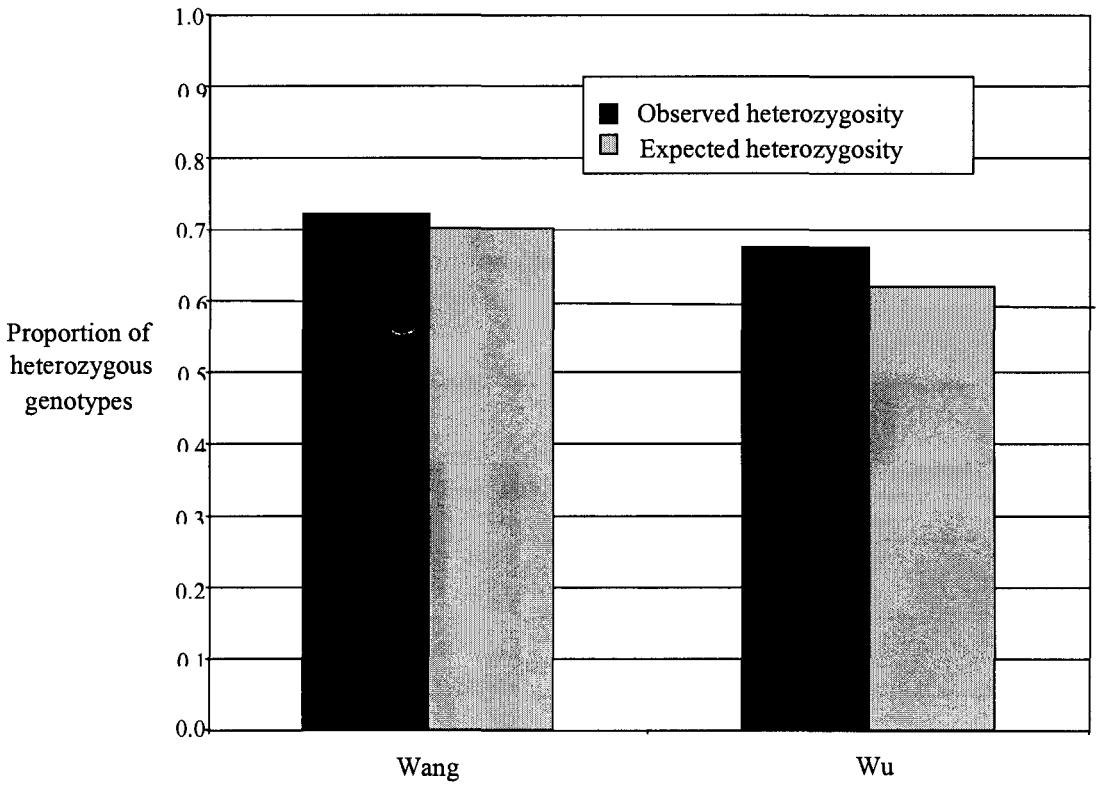
6.4 Expected and observed heterozygosity

In both the Wang and Wu pedigrees expected heterozygosity was less than observed heterozygosity, opposite to the results obtained from the unrelated Hui sample populations. For the Wang pedigree, average expected heterozygosity was calculated as 0.703 while the average observed heterozygosity was 0.722. In the Wu pedigree, the average expected and observed heterozygosity values were 0.619 and 0.674 respectively (Table 6.1, Figure 6.7). The comparable figures for the unrelated Hui sample population were an average expected heterozygosity of 0.755 and an average observed heterozygosity of 0.457 (Table 5.1, Figure 5.1).

The level of expected heterozygosity in the Wang was not significantly different to the random sample population ($p = 0.185$). By comparison, the Wu pedigree was significantly different to the random sample population ($p = 0.017$). Furthermore, in both the Wang and Wu pedigrees observed heterozygosity was significantly larger than in the random sample population ($p = 0.001$ and $p = 0.010$ respectively).

The overall pattern of expected and observed heterozygosity levels was surprising given the limited number of alleles available. However higher than expected levels of heterozygosity also have been observed in other communities with a high prevalence of consanguineous unions, and the conclusion reached was that some form of selection against homozygosity may be operating at early development gene loci (Wang *et al.* 2000). As the Hui also permit and practise consanguineous unions, a similar phenomenon may be operating in these pedigrees. However, once again, the sample sizes may be inadequate to permit unbiased statistical analysis.

Figure 6.1 Observed and expected heterozygosity in the Wang and Wu pedigrees



6.5 Allelic correlation coefficient

As with the Hui and Han unrelated sample populations, *f* values were calculated for both the Wang and the Wu as a measure of deviation of genotypes from HWE. It was found that the *f* value for the Wang was –0.062 and for the Wu it was -0.140. This is sharp contrast to the random sample population, which recorded an *f* value of +0.401. These negative *f* values indicate an avoidance of inbreeding (Section 3.3.5), in contradiction to the analysis of pedigree structure and calculation of *F* values (Section 6.2) which indicated the presence of consanguinity. The smaller size of the Wu pedigree suggests that the computed value may not be statistically robust, but the larger sample size of the Wang also shows a negative correlation coefficient. On the assumption that non-paternity

does not occur in the two kindreds, a convention supported by the inheritance patterns of the microsatellite markers, these results in concert with section 6.2 are suggestive of some form of selection against homozygote genotypes.

Figure 6.2 Correlation coefficients calculated from the Wang pedigree

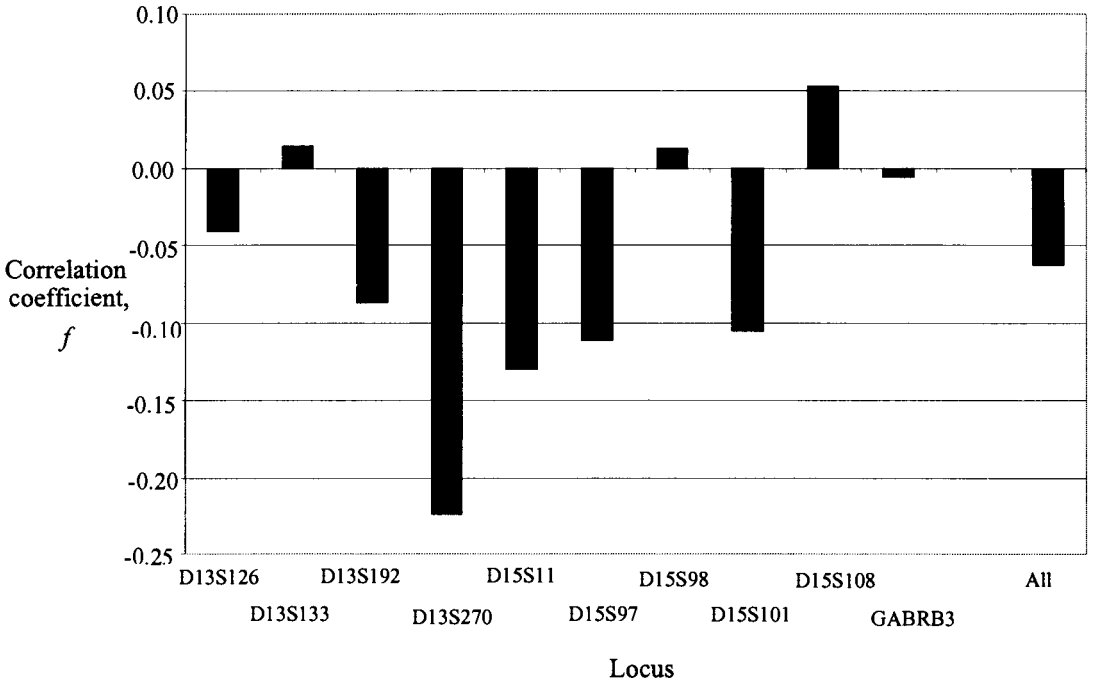
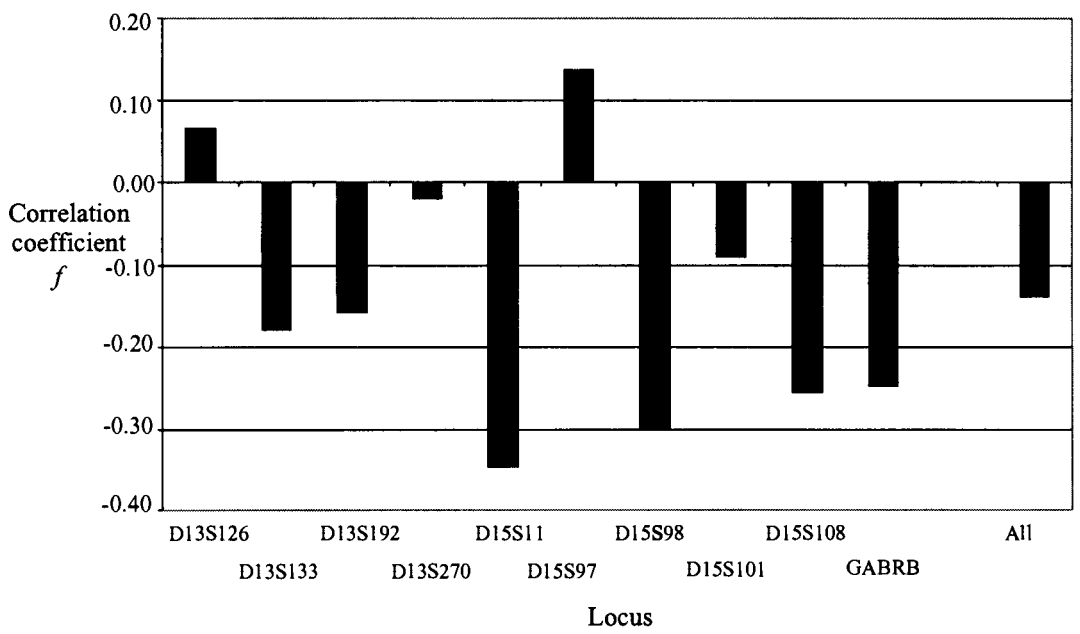


Figure 6.3 Correlation coefficients calculated from the Wu pedigree



6.6 Polymorphic information content

Given the paradox of high heterozygosity and low polymorphism found in the pedigree data, the polymorphic information content, PIC, also was calculated to assess of the effect of consanguinity. The PIC value is the probability that the genotype of a given offspring will permit identification of which marker allele at a locus was inherited from each parent. Thus, in effect PIC measures the potential of a marker for use in identity testing. The formula is as follows:

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 p_i^2 p_j^2$$

Equation 6.1 Polymorphic information content

Figures 6.3 and 6.4 show that the computed PIC values were well below the observed heterozygosity values for both pedigrees. Average PIC for the Wang was 0.587 and for the Wu 0.499, while average observed heterozygosity was 0.775 and 0.658 respectively. Therefore the PIC values reflect the reduced allele numbers found in both pedigrees.

Like heterozygosity, PIC decreases in proportion to the number of alleles. However PIC provides a more conservative estimate of heterozygosity from the observed allele frequencies (Taylor *et al.* 1994). If a PIC is low, it indicates a high probability that the parental haplotypes are identical with a corresponding increased potential for the homozygous genotype. Under these circumstances PIC would be expected to indicate a greater deviation of the observed heterozygosity from expected heterozygosity. Thus it may be a better indicator than observed heterozygosity if there is selection against homozygous.

In summary, the combined results from the analysis of observed heterozygosity and calculation of correlation coefficients for the two pedigrees support a hypothesis of heterozygote advantage. The PIC values of the two pedigrees are smaller than the observed heterozygosity levels indicating that selection against homozygous genotypes may be occurring. If such selection was not apparent the PIC values would have been higher, more closely resembling the level of observed heterozygosity.

Figure 6.4 Comparison of PIC, observed and expected heterozygosity levels in the Wang pedigree

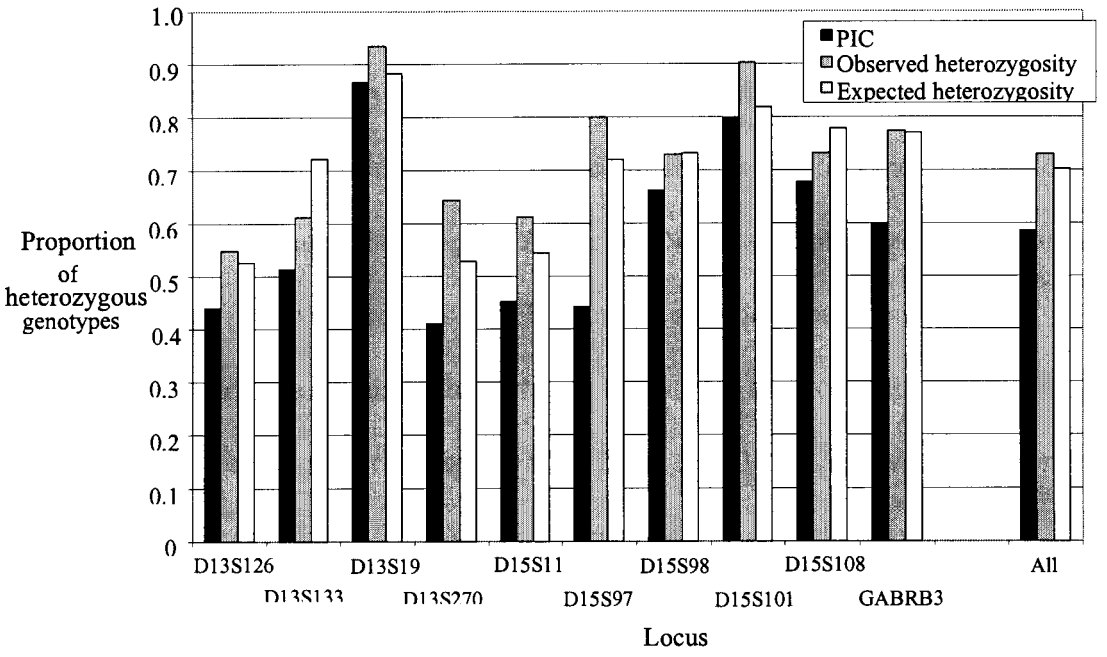
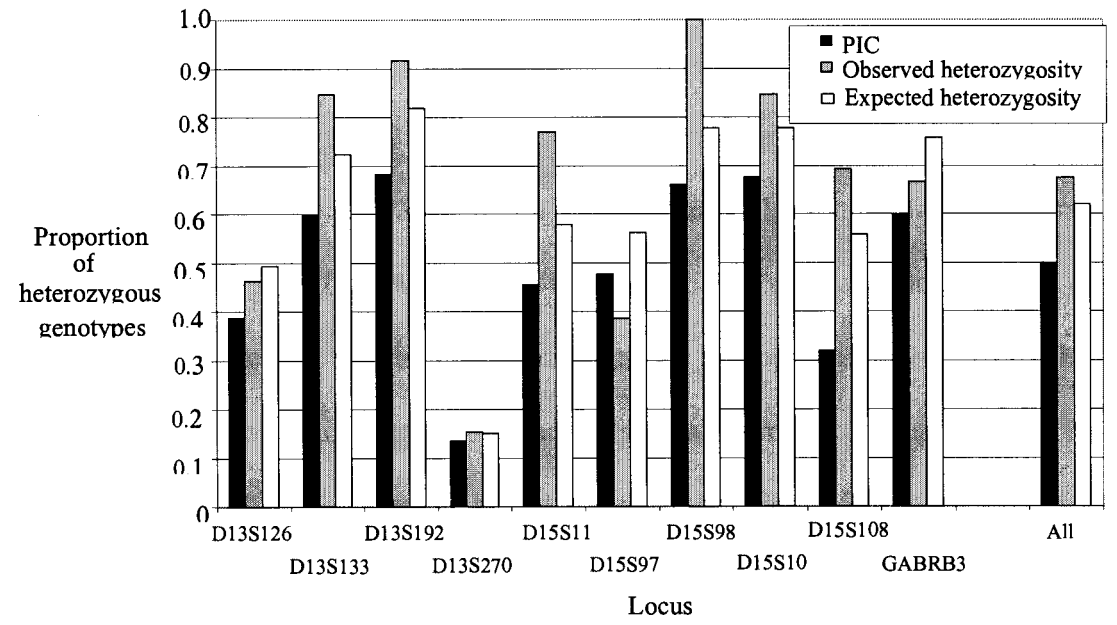


Figure 6.5 Comparison of PIC, observed and expected heterozygosity levels in the Wu pedigree



6.7 Y-chromosome allele distributions

A total of 16 Y-chromosome haplotypes were extracted from the Wang pedigree and 6 haplotypes from the Wu pedigree (Appendix C). The haplotypes obtained from the pedigrees were first compared to the Hui random sample population via analysis using the ARLEQUIN software package. No Wang or Wu haplotypes were shared with the Hui random sample population.

Further analysis of the pedigree haplotypes showed that the Wu pedigree had only one allele for markers DYS392 and DYS389I with two alleles for each of the other markers. The Wang pedigree had nearly triple the number of haplotypes observed in the Wu pedigree, but only approximately double the number of alleles per locus (Table 6.2).

Calculation of the mean number of alleles showed that the Wang pedigree had an average of 3.29 alleles per locus and the Wu had 1.17 alleles per locus. The allele size variance was equivalently low with the Wang exhibiting a variance of 2.43 and the Wu 0.93. Statistical assessment showed that the mean number of alleles per locus in the Wang was not significantly different to the random sample population ($p = 0.095$) whereas the mean number of alleles per locus in the Wu were significantly smaller that the random sample population ($p = 0.01$). The latter finding is mostly probably due to the small number of samples in the Wu pedigree, adversely affecting the reliability of the testing. The same effect was seen when allele size variance was compared with the random sample population: the Wang ASV was not significantly different ($p = 0.425$) but the Wu ASV was significantly smaller ($p = 0.021$).

As hypothesised when the autosomal data were analysed, the similarity between the ASV and mean number of alleles per loci in the Wang and the corresponding values in the random Hui sample population can be interpreted as evidence of residual diversity originating from the diverse founding populations. If this is correct, then either the level of diversity has not been greatly reduced by a history of consanguineous marriage, or close kin marriage may be a recent or intermittent practice.

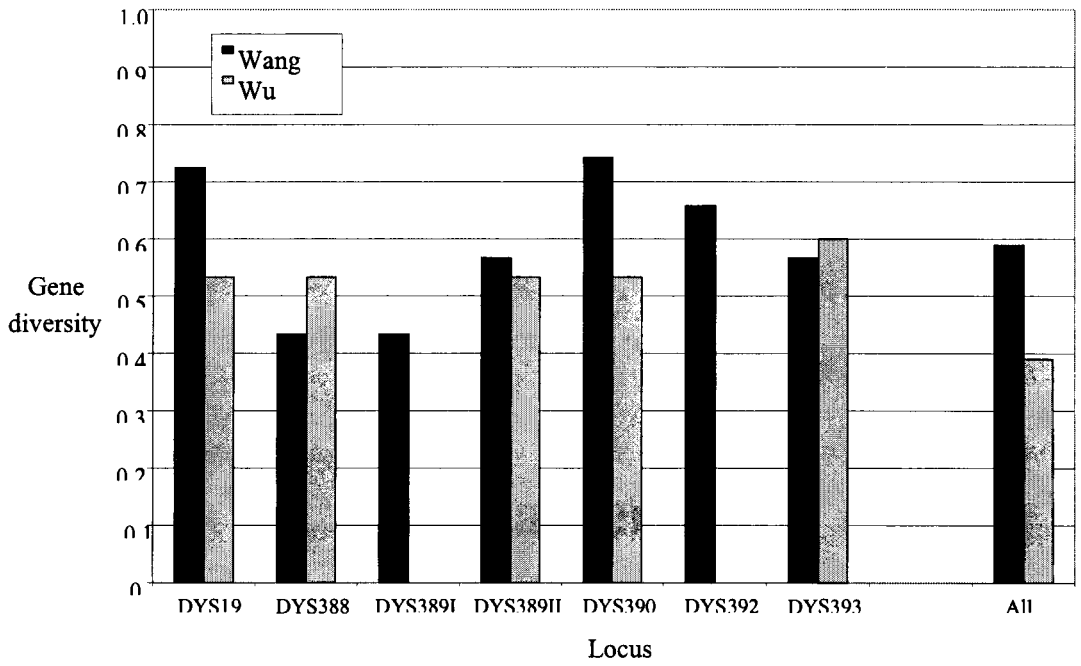
The average level of gene diversity in the Wang pedigree was 0.589, and 0.390 in the Wu pedigree (see Figure 6.6). Calculation of statistical significance showed that the level of gene diversity in the Wang did not differ significantly from random sample population ($p = 0.148$), but the level of gene diversity in the Wu was significantly smaller than the random sample population ($p = 0.04$). As with the comparison of ASV and mean number of alleles, the significant result for the Wu is most probably associated with the small available sample size. Meanwhile, the non-significant result for the Wang also mirrors the comparisons of ASV and mean number of alleles, as this relatively high level of gene diversity is a further indication of residual diversity originating from the diverse founding populations.

Table 6.2 Summary of allele distributions of Y-chromosome microsatellite markers in the Wang and Wu pedigrees

Marker	Population	Allele range (bp)	No. of alleles	MFA*	ASV**	Gene Diversity
DYS19	Wang	182-202	4	202	4.3	0.725
	Wu	190-194	2	190	0.5	0.533
DYS388	Wang	126-135	3	128	2.3	0.433
	Wu	129-138	2	137	4.5	0.533
DYS389I	Wang	249-257	3	249	1.0	0.433
	Wu	253	1	253	0.0	0.000
DYS389II	Wang	361-377	3	361	4.3	0.567
	Wu	369-373	2	369	0.5	0.533
DYS390	Wang	207-219	4	211	1.7	0.742
	Wu	215-219	2	215	0.5	0.533
DYS392	Wang	248-260	4	245	2.9	0.658
	Wu	245	1	257	0	0.000
DYS393	Wang	115-119	2	115	0.5	0.567
	Wu	115-119	2	115,119	0.5	0.600
Average	Wang	--	3.29	--	2.43	0.590
	Wu	--	1.71	--	0.93	0.390

* MFA = Most frequent allele
** ASV = Allele size variance (variance of allele repeat size)

Figure 6.6 Y-chromosome gene diversity in the Wang and Wu pedigrees



Chapter 7

Discussion

7.1 Introduction

Recent advances in genome technology, such as the polymerase chain reaction and the development of fluorescent DNA sequencing, have enabled researchers to qualitatively and quantitatively assess human genetic variation in concert with demographic, historical and anthropological data. For example, in studies of the genetic anthropology of Amerindians, where at least three different migration events have been traced by the combination of anthropological and genetic research (Karafet *et al.* 1999).

The main purpose of the present study was to investigate correlations between genetic variation and human history, focusing on the Han and Hui peoples of Liaoning province, PR China. The history of these peoples is essentially that of population migration, amalgamation and then, most predominantly in the Hui, reproductive isolation. In the following discussion, it will be seen that through the combined study of historical and genetic information on the Han and the Hui populations, an overview of the relationship between intra- and inter-population genetic diversity may be formulated.

7.2 Patterns of genetic diversity within the Han and Hui

It was evident from the analysis of the random sample populations of the Han and the Hui that the genetic characteristics of both ethnic groups are primarily defined by genetic differences within each population. This conclusion can be drawn by comparing the results of the population differentiation exact tests with those obtained from AMOVA analysis of autosomal data (Tables 5.1 and 5.7). The exact test results exhibited significant differences between the populations at 8 of 10 loci.

Yet AMOVA results produced only a small Φ_{ST} value for the pooled data set from the two chromosomes studied, chromosomes 13 and 15, and when the data for the two chromosomes were analysed separately, both yielded non-significant Φ_{ST} values (Table 5.4). Thus, the low Φ_{ST} values indicate that the strong genetic differentiation between the Han and Hui shown by the exact tests is predominately due to their different internal population structures.

While the analysis of autosomal alleles revealed that intra-population diversity was the main factor determining differentiation between the Han and the Hui, the results of the Y-chromosome analysis produced a different conclusion. A comparison of the exact test and AMOVA results for the Y-chromosome data indicated a more significant role for inter-population diversity in the relationship between the Han and the Hui. The exact tests for population differentiation showed that 4 of the 7 loci exhibited significant differentiation between the Han and the Hui (Table 5.6). In contrast, the AMOVA results indicated a much larger between-population variance, accounting for 13.99% of total variation (Table 5.7). Thus, the Y-chromosome haplotype distribution appears to be more affected by inter-population diversity than autosomal genotypes, and so it is better suited to inter-population genetic diversity analysis. This topic is further discussed in Section 7.3.

In summary, the population structures of the Han and the Hui most probably result from different historical backgrounds and differing cultural practices. Consequently, the discussion of genetic diversity within the Han and Hui is considered separately.

7.2.1 Genetic diversity in the Han

The term Han *minzu* is a politically convenient label to encompass the vast majority of the Chinese population. It was initially used at the beginning of the 20th century to develop a sense of national unity and thus facilitate the fall of the Qing dynasty, which occurred in 1908 (see Sections 2.2 and 2.6.5). The wide allele size distributions and the diverse range of Y-chromosome haplotypes revealed by genetic analysis of the Han of Liaoning province shows that this sense of Han *minzu* is of limited value in purely genetic terms.

This hypothesis is backed up by available historical information. For example, Du and Yip (1993) describe the Han as an ethnic group based on the ancient Huaxia of the middle and lower reaches of the Yellow River. It is believed that the Han of the north east provinces of China originated from groups in the south of the country who settled in the North East in order to solidify dynastic control of the region (Fairbank and Reischauer 1990). Through generations, the Han would have intermingled with the indigenous Jurchen tribes (now known as the Man or Manchu *minzu*), and the Mongolian peoples of the North East region, resulting in a heterogenous population.

The patterns of genetic diversity found in the Han of Liaoning supports this notion of a population with a heterogenous mix of different ethnic origins. The random sample of the Han exhibited a large array of alleles, averaging approximately eleven alleles per locus (Table 5.1). Nonetheless, the observed level of heterozygosity in the Han was significantly lower than in the reference Caucasian population. The combination of a large array of alleles but a restricted array of genotypes may be interpreted as an indication of the existence of population stratification (Section 5.9). In addition, a variety of Y-chromosome

haplotypes (Section 5.12) and a high level of Y-chromosome gene diversity, were found (Section 5.11), adding further support to the hypothesis of population stratification.

To fully appreciate the population structure of the Han of Liaoning, it would have been helpful to have access to greater anthropological detail on their population structure and origins, including extended pedigrees for genetic analysis. This was not possible as the sampling procedure adopted was originally focused on the study of genetic differentiation and population structure in the Hui, using a random sample Han population as the local reference population.

In conclusion, further anthropological research into the Han of Liaoning province is needed to provide a more comprehensive analysis of their genetic structure. It is therefore proposed that, if possible, further demographic and genetic studies should be undertaken to elucidate the source, type, and nature of the variation and structure of the Han population.

7.2.2 Genetic diversity in the Hui

To fully understand the patterns of genetic diversity in the Hui of Liaoning, it is appropriate to note the relatively small size of the Hui population, i.e., approximately 263,000 individuals, by comparison with 33 million Han who are resident in the province (Family Planning Commission 1997). Therefore, when collecting a random sample population, the probability of inadvertently selecting related individuals would be much higher in the Hui than the Han. The combined effect of restricted local population sizes and endogamous marriage patterns could explain the statistically significant heterozygote deficiency demonstrated by HWE analysis (Table 5.2b), and the positive average correlation coefficient of 0.400 (Figure 5.2).

Contrasting results were found in the analysis of the Wang and Wu pedigrees. In summary, an excess of heterozygous genotypes was observed in both pedigrees, demonstrated by the average correlation coefficient value of –0.062 for the Wang and –0.140 for the Wu (Figures 6.2 and 6.3). This is unexpected, as study of the pedigrees indicates the prevalence of consanguineous marriages, and population genetics theory suggests that one result of inbreeding would be an increase in homozygosity (Hartl and Clark 1997 pp 135 - 149).

One explanation for the paradox between the low level of heterozygosity in the Hui random sample population and the high level of heterozygosity in the two Hui pedigrees may be the existence of groupings within the wider Hui community based on cultural and religious affiliations. It has been shown that many different Islamic sects exist within the Hui *minzu* (see chapter 2). While divisions of this type are relatively recent in the history of the Hui people (200-300 years), there could have been a sufficient number of generations of restricted marriage within specific religious sub-sects to initiate the formation of smaller reproductively isolated populations in the Hui population of Liaoning.

Groupings could also be based on the urban/rural divide, as urban Hui communities tend to differ from their rural counterparts in terms of adherence to religious and cultural practices (Gladney 1998). In any event, such population stratification would be evident, in a genetic sense, by the presence of lower than expected levels of hetrozygosity in the wider Hui population. Finally, the enormous loss of life experienced in the wars of the mid-19th century (Section 2.6.5), during the Japanese invasion and the subsequent Civil War during the 1930s and 1940s, and later by famine in the 1960s, all could have exerted a profound effect on the genetic structure of the various Hui communities in China.

Of course, the observed contrasts might also be due to the effect of small sample size. Only 53 Hui samples were available for the random sample population analysis, and there were even fewer pedigree samples, 31 Wang and 14 Wu. In addition, only 10 autosomal loci from two chromosomes were analysed, along with 7 Y-chromosome loci. The analysis of this relatively small data pool may produce results that do not reflect the true nature of the Hui gene pool.

For a more conclusive analysis of genetic diversity in the Hui, further samples would need to be taken. To establish statistically robust results, it would also be beneficial to test a larger number of loci from a wider range of chromosomes. This proposed extended analysis of the population genetics of the Hui would however also require knowledge of any population stratification revealed through demographic and anthropological studies.

7.3 Comparison of the autosomal and Y-chromosome genetic diversity of the Han and Hui

According to the historical record two of the major founding populations of the present Hui population were Arab males and Han females (Wong and Dajani 1988). Therefore, it would be expected that the distribution of autosomal alleles in the modern Han and Hui might exhibit similarity due to the shared female ancestry.

The comparison of autosomal and Y-chromosome AMOVA results seem to confirm with this hypothesis. To put the AMOVA results into perspective, it has been found that the global population variance for autosomal STRs is $\Phi_{ST} = 0.15$ (Barbujani 1997). The autosomal between-population proportion of total variance obtained for the Han and Hui was just 4.63% ($\Phi_{ST} \cong 0.046$), indicating a

much closer genetic relationship than found between the major continental human populations such as European, Amerindian, Southern Asian and Northern Asian. The most obvious cause is the initial admixture of Han females in the creation of the Hui population, and possible subsequent mingling of the populations. Meanwhile the Y-chromosome haplotype between-population proportion of total variation was 13.99% ($\Phi_{ST} \cong 0.1399$) (Table 5.7), a value closer to the worldwide autosomal average. In addition, this proportion of variance is higher than corresponding values calculated from a comparison of European populations (de Knijff *et al.* 1997) (Section 5.12).

However, as stated in Section 5.12, the direct comparison of autosomal and Y-chromosome Φ_{ST} values may not be justified. Firstly, the autosomal Φ_{ST} parameter is a measure of genotypic between-population variance, while the Y-chromosome Φ_{ST} is a measure of haplotype between-population variance. This difference is due to the different ways in which autosomal and Y-chromosome markers are inherited. The Y-chromosome is effectively haploid and is passed from father to son, while autosomal markers are diploid and transmitted by both parents to sons and daughters. This difference leads to an effective Y-chromosome population that is approximately one quarter the size of the effective autosomal population (Péréz-Lezuan *et al.* 1997b). In addition, the diploid nature of the autosomal markers leads to autosomal diversity being influenced by recombination during meiosis. In summary, the lower autosomal Φ_{ST} may not be due solely due to female admixture between the Han and the Hui, but also possibly to the larger effective population size and the effect of recombination.

The analysis of the hypervariable regions (HVR-I and HVR-II) of the human mtDNA genome could provide a direct assessment of female gene flow free from the complexities of recombination. The mtDNA genome is maternally inherited and therefore haploid. Hence, mtDNA diversity, like Y-chromosome diversity, is measured on a haplotypic basis, allowing for a more direct comparison of the female and male gene pools of the Han and the Hui. A comparison of Y-chromosome haplotypic diversity and mtDNA haplotypic diversity would be more likely to confirm the presence of any differential patterns of male and female admixture and migration (Comas *et al.* 1998).

7.4 Y-chromosome microsatellites and male gene flow in the Hui

It was demonstrated in the previous section that the male Han and Hui show significantly different distributions of Y-chromosome haplotypes. Further analysis of male Hui genetic diversity, in combination with the analysis of available historical information, demonstrates the existence of diverse ancestral origins in the male Hui population.

The Hui had a relatively high average Y-chromosome gene diversity of 0.656, compared with the values from other studies in which the same markers were analysed (Section 5.11). This is indicative of an array of Y-chromosome haplotypes and hence diverse male ancestry. Further evidence of this diversity was seen in the relatively large pairwise difference in allele numbers between the Hui haplotypes, with an average difference between Hui haplotypes of approximately 5 alleles for the seven locus haplotypes, and 4 for the six locus haplotype. In addition, the analysis of pedigree Y-chromosome haplotypes showed that they all differed from the random sample population haplotypes (Section 6.7). Therefore, the actual level of haplotypic diversity in the male Hui

of Liaoning may be even greater than indicated by the analysis of the random sample population.

The next step in confirming the existence of diverse male ancestry in the Hui was by tracing gene flow through phylogenetic analysis. The construction of unrooted neighbour-joining trees exhibited a consistent pattern of greater similarity between male Hui and male Han than with other populations, such as Southern Chinese, Mongolian and Central Asian populations. However, it was previously shown that the Y-chromosome haplotype Φ_{ST} value for the male Han and the male Hui populations was 0.1399, a large value compared to Φ_{ST} values found from comparisons of European populations, which ranged from 0.007 to 0.0812 (de Knjiff *et al.* 1997). It could be concluded that, from the perspective of phylogenetic analysis, microsatellite markers are efficacious in clarifying genetic divisions between closely related populations. But as the genetic relationship between population decreases, they become less useful in resolving genetic relationships.

It seems, however, that a major factor confounding the present phylogenetic analysis stems from the founding history of the Hui, which involved the amalgamation of many different populations. For example, the available historical information suggests origins stemming from Turkic, Iranian and Arab populations, which themselves may have been genetically heterogenous (Section 2.6.1). The large range of alleles present in the male Hui population may result from population admixture caused by such population movements. If so, tracing male gene flow using just microsatellit. markers would be very difficult, as the distribution of alleles would have become increasingly heterogenous due to the intermingling of these populations. Therefore, the analysis of male Hui gene flow

may be best served by tracing the various genetic origins separately, rather than by attempting a broad overview of male Hui genetic diversity as a whole.

In attempting to trace the different lineages of the male Hui of Liaoning, it may be more advantageous to analyse the distribution of less polymorphic markers, such as SNPs. For example, a recent study used SNPs to trace the origins of Amerindian tribes (Karafet *et al.* 1999). This resulted in the discovery of several founder haplotypes from which ancestry was traced to a region surrounding Lake Baikal, which anthropologists previously had suggested as the possible origin of the ancestors of Native Americans.

Conceptually, SNPs can be visualised as defining the fundamental branches of a phylogenetic tree, while more polymorphic loci, such as microsatellites, would be used to recognise the finer divisions. A more complete phylogeny of the male Hui population could thus be uncovered by the combined use of a haplogrouping strategy involving both SNPs and microsatellite loci.

7.5 Consanguinity and genetic diversity

The final part of the investigation was assessment of the effect of consanguinity on the level of genetic diversity in the Hui. To date, the major focus of studies into human consanguinity has been the increased risk of morbidity and mortality, due to the higher probability of the expression of autosomal recessive disorders. In general, it has been concluded that the less common the disorder, the greater the influence of consanguinity on its prevalence (Bittles 1998). However, there has been little study of the effect of consanguinity on genetic diversity in ethnic groups.

In the present study, the genetic structure of two Hui pedigrees, the Wang and the Wu, was analysed to specifically investigate this effect. There was a

modest prevalence of consanguineous marriage in both the Wang and Wu indicated by estimated mean pedigree inbreeding coefficients of $\alpha = 0.003$ and $\alpha = 0.007$ for the Wang and Wu respectively (Section 6.2). However, genetic analysis of both pedigrees indicated the presence of excess heterozygosity, demonstrated by their negative genotypic correlation coefficients of –0.062 and –0.140 for the Wang and Wu respectively (Figures 6.2 and 6.3).

As it has been shown in the population study, the Hui have diverse backgrounds demonstrated by the existence of wide allele distribution. One possible reason for these findings is that consanguineous marriage had not been continuously practised over a sufficiently long number of generations in the Hui community, and at an appropriately high prevalence, to reduce the number of alleles present in the pedigrees.

Another possibility is associated with the effect of small sample size and sampling error. The two pedigrees were of different sizes; the Wang encompasses 81 individuals over 5 generations while the Wu pedigree includes just 17 individuals over 3 generations. In addition, only 34 of the 81 individuals could be sampled from the Wang and only 14 from the Wu. Since just 10 loci from 2 chromosomes 13 and 15 were surveyed, error due to the small number of samples and number of loci surveyed may have arisen.

However, other studies being undertaken in the Centre for Human Genetics, Edith Cowan University, suggest that the excess heterozygosity observed is not due to sampling error. The same markers used in the present study have been analysed in pedigrees from Pakistan and Southern India, populations in which consanguineous marriage is commonplace.

In both sets of analyses, significant heterozygosity excess was found. Since the studies involved much larger pedigrees and, in Southern India, the analysis of many more markers, they serve to confirm the veracity of the findings obtained in the present study.

Therefore, in light of these results, a third possibility must be considered, that is the occurrence of selection. The low PIC values calculated for the Wang and Wu (Section 6.6) in combination with the high observed heterozygosity, indicate the existence of selective processes. These PIC values basically suggest a high probability that parental haplotypes in the pedigree are identical, therefore a greater potential for the existence of homozygous genotypes in their offspring would be expected. However, the observed heterozygosity levels of 0.772 and 0.674 for the Wang and the Wu suggests either that the potential for the formation of homozygous genotypes is not realised, or more probably, there is early prenatal selection against the survival of homozygous genotypes.

This hypothesis is supported by reference to other human and animal studies. For example, studies by Neel and Ward (1972), on the Yanomama Amerindians, semi-nomadic people who live on the Brazilian-Venezualan border, and Workman *et al.* (1973) who studied the Papago of Arizona, U.S.A., both showed negative correlation coefficients, that is excess heterozygosity, observed in populations with a relatively high level of inbreeding. It was suggested in both studies that a difference in gene frequencies between the sexes was an explanation for the presence of excess heterozygosity. However, a more recent study on an inbred Amerindian tribe in Arizona, the Havasupi, attributed a similar excess of heterozygotes at the HLA-A locus to balancing selection, that is, to heterozygote advantage (Markow *et al.* 1993).

A review of various animal studies shows stronger evidence for selection as a cause of excess heterozygosity in inbred populations. For example, a long-term study into the effects of inbreeding in a herd of Speke's Gazelle at the St Louis Zoo, USA (Templeton and Read 1983, 1984, 1998) demonstrated an initial adaptation to high levels of inbreeding, with the eventual elimination of inbreeding depression. The mean coefficient of inbreeding calculated from the pedigree was $\alpha = 0.149$, yet an isozyme study showed an allelic correlation coefficient of $f = -0.291$ in the same animals. The high α value arose because the herd was founded by just three females and one male, whereas it was concluded that the negative f value was due to management of the herd to minimise inbreeding, i.e., to artificial selection processes.

More recently, studies on inbreeding and heterozygosity have focused on microsatellite markers. In both harbour seal pups (Coltman *et al.* 1998) and red deer (Coulson *et al.* 1998), a positive correlation was demonstrated between microsatellite heterozygosity and neonatal survival, with the heterozygous animals more likely to reach reproductive age. Therefore, microsatellite heterozygosity was positively linked to overall biological fitness. This result is problematic as it is generally accepted that microsatellites play no functional role in the genome. As a result they would have a constant rate of evolution independent of the size of a population and as a consequence, are selectively neutral (Charlesworth *et al.* 1994, Jobling 1995, de Knijff *et al.* 1997).

A solution to this dilemma was presented in a study of the population genetics of the oyster, *Ostrea edulis.* It was shown that microsatellite multi-locus heterozygosity enhanced the early developmental growth of oyster larvae. The suggested mechanism was a "hitchhiking effect" in which over-dominant

heterozygous genotypes were linked to fitness-associated genes affecting early developmental growth, which in turn, were subject to selection.

In conclusion, excess microsatellite heterozygosity may be indicative of a compensatory mechanism for the retention of biological fitness in populations that have experienced long-term inbreeding. Therefore, a simple positive correlation between consanguinity and reduced heterozygosity, as predicted by classical population genetics theory, may not be entirely correct. Rather, there may be a more complicated scenario involving both the history and the types of consanguineous marriage in the ethnic group.

7.6 Overall conclusions

The major difficulty in studying human population genetics is that, due to the impact of cultural influences on the population dynamics of human populations, humans are the organisms least likely to conform to the expectations of population genetics models (Jorde 1997). However, by treating historical and anthropological aspects of the Han and Hui populations as variables that effect their overall genetic variation, a broad survey of the population genetics of the Han and Hui is possible.

Unfortunately, only limited historical and anthropological information is available on the study populations, and so the nature of the population genetics of the Han and the Hui could only be cautiously inferred. For a more detailed analysis of the genetic structure of the Hui and Han communities, more substantial anthropological and demographic data on the social and cultural structures of the populations is needed. Thereby a more complex hierarchy of genetic variance could be established that more closely reflected the actual structure of each population.

More detailed anthropological studies would also allow the construction of better defined patterns of male gene flow within the Hui, and result in an improved understanding of the possible candidate founding populations. The simultaneous analysis of biallelic and polymorphic Y-chromosome markers would be the appropriate genetic methodology for a more precise phylogeny of male gene flow in the Hui. This technique has proven to be especially useful in elucidating the origins and migration patterns of Amerindians (Karafet *et al.* 1999), and the migration of early humans out of Africa (Hammer *et al.* 1998). It is therefore recommended that Y-chromosome biallelic markers should be surveyed to genetically define forbears of the male Hui, an ancestry that, according to historical sources, can be traced back to a combination of Central Asian, Turkic, Iranian and Arab founder populations.

While this study is based on the present anthropology and population genetics of two Chinese populations, the genetic heritage of both populations may well reflect populations that have long since ceased to exist in a formal manner, and are only known from historical record. An example is the possible migration of Caucasians into North West China approximately 4000 years ago, and their continued existence into the beginning of the modern era.

Evidence of this migration is present today in the form of mummified remains in the museums of the city of Ürümchi in Xingjiang province, P.R. China. The local Muslim Uygur community of Xingjiang autonomous region consider these mummies to be evidence of their ancient ancestors. This belief may be valid, as many non-Chinese people now living in this province have what often are considered Caucasian physical features, such as blue eyes and red hair, possibly resulting from the genetic heritage transmitted by their forebears

(Wayland Barber 1999). Thus great care must be taken in relating one Chinese population to another, as there may have been intermixture with now formally extinct populations that contributes to current population genetic structure.

In conclusion, to take account of the complexities of population stratification, sex-biased ancestry and the possible inheritance of alleles from extinct populations, future studies into the genetic variation of the Han and Hui should involve the simultaneous analysis of several different genetic systems. For example, differences in the structure of male and female genetic diversity within and between populations could best be investigated by comparisons of mtDNA and Y-chromosome haplotypes, while autosomal genotype analysis is better suited for more general analyses of intra-population structure.

An additional layer of complexity is added to the pattern of population genetic structure by preferential consanguineous marriage. From the results of the present study into the genetic structure of the Hui of Liaoning province, consanguinity has been shown to exert no direct influence on molecular diversity at microsatellite loci. Rather, consanguinity is more likely to be a factor in the generation of population stratification. Further studies involving the investigation of population sub-structure in the Hui would be required to validate this hypothesis.

The overall effect of consanguinity is, however, important in terms of public health. Many studies have correlated consanguinity with increased risk of stillbirths, infant deaths and autosomal recessive diseases (Bittles and Neel 1994, Dorsten *et al.* 1999, Stoltenberg *et al.* 1999). The compensatory effect of early prenatal selection hypothesised in this study may reduce these risks in populations where consanguineous marriage has been practised for many generations. The

possible existence of selection in the inheritance of early development genes in humans, as already hypothesised in animal studies, demonstrates that selection may be an important factor in the health of inbred populations. Therefore, further investigation into the patterns of expression of early development genes is required.

Addendum

In addition to the Han and Hui sample populations, the Centre of Human Genetics, Edith Cowan University, has obtained samples from three other Muslim ethnic groups (Bonan, Sala and Dongxing) and three non-Muslim ethnic groups (Tibetan, Yaozu and Baizu) from P.R. China. Therefore the current investigation can be considered as part of a wider investigation undertaken by Dr Wei Wang and Prof. Alan Bittles of the Centre for Human Genetics, Edith Cowan University, into the genetic structure of ten unrelated and geographically dispersed populations resident in the Peoples Republic of China (PR China). In this larger study, the genomic information obtained will be compared with blood group and enzyme polymorphism data, and anthropological and historical sources, to determine whether parallels between genetic diversity and historical detail can be sustained.

Appendix A

Publications and useful web sites

A.1 Conference abstract presented at HGM 1999, HUGO, March 1999, Brisbane QLD.

A comparison of autosomal and Y-chromosome allele profiles in co-resident Han and Hui communities in northeast China

W. Wang[1], M. Black[1], H. L. Jia[2], Cong Qian[3],and A.H. Bittles[1]
[1]Edith Cowan University, Perth, Australia, [2]P.L.A. No. 201 Hospital, Liaoyang, PR China and [3]China Medical University, Shenyang, P.R. China

Chinese historical records indicate that the Hui, a Muslim minority population which currently numbers over 8 million, originated from the marriage of non-Chinese males with Han females. A sizeable proportion of the Hui male forebears are believed to have been Muslim traders from the Middle East, Iran and Central Asia who were involved in the Silk Road, which operated between Xi'an in China and Constantinople/Istanbul on the Bosphorus from approximately 120 BC to 1600 AD.

To investigate the degree of genetic admixture between co-resident Hui and Han communities, finger-prick blood samples were collected from randomly selected individuals, 100 Han males and 37 Hui males and 36 females, in Liaoning province, northeast China.. DNA was obtained by chloroform-phenol extraction and the samples analysed on an ABI 373 DNA Sequencer. Ten dinucleotide markers on chromosomes 13 and 15, and 7 tri- and tetranucleotide markers on the Y-chromosome, were run on samples from both communities. The Hui demonstrates a quite different genetic profile on the 10 autosomal markers from that of the co-resident Han. Nine out of 10 autosomal makers show significant deviation from the Hardy-Weinberg equilibrium when testing the Hui samples for heterozygote deficit, which is well correspondent to the wide practice of preferential marriages in the same ethnic group and even within same family names in the Hui. The investigation of the genetic profiles of the Y-chromosome markers among Hui and Han, which might provide information on the male-mediated gene flow in the Hui community, is under progress. The initial results obtained with the Y-chromosome markers also indicate significant disparity between the Hui and Han communities, with less diversity in the Hui in terms of allele numbers and allele ranges.

A.2 DNA Polymorphism Vol 8 2000, Tokyo, Japan. (in press).

Autosomal and Y-chromosome allele profiles in co-resident Han and Hui communities in northeast China.

Wei Wang[1], Michael Black[1], Cong Qian[2], Huling Jia[3] and Alan Bittles[1]
[1]Centre For Human Genetics, Edith Cowan University, Perth, Australia, [2]China Medical University, Shenyang, P.R. China and [3]P.L.A. No. 201 Hospital, Liaoyang, P.R. China

A.3 Useful web pages

Arlequin
Home page: http://anthropologie.unige.ch/arlequin/

Centre d'Etudes du polymorphisme Humain (CEPH)
Home page: http://www.cephb.fr

Eurasia 98
Home page: http://www.imm.ox.ac.uk/~eurasia/htdocs/index.html

Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University
Home page: http://ruly70.medfac.leidenuniv.nl/~fldo/

Genepop
Home page: http://www.cefe.cnrs-mop.fr

Genome Database
Home Page: http://www.gdb.org

Human Genome Diversity Database
Home page: http://human.stanford.edu/

Microsat
Home page: http://human.stanford.edu/microsat/microsat.html

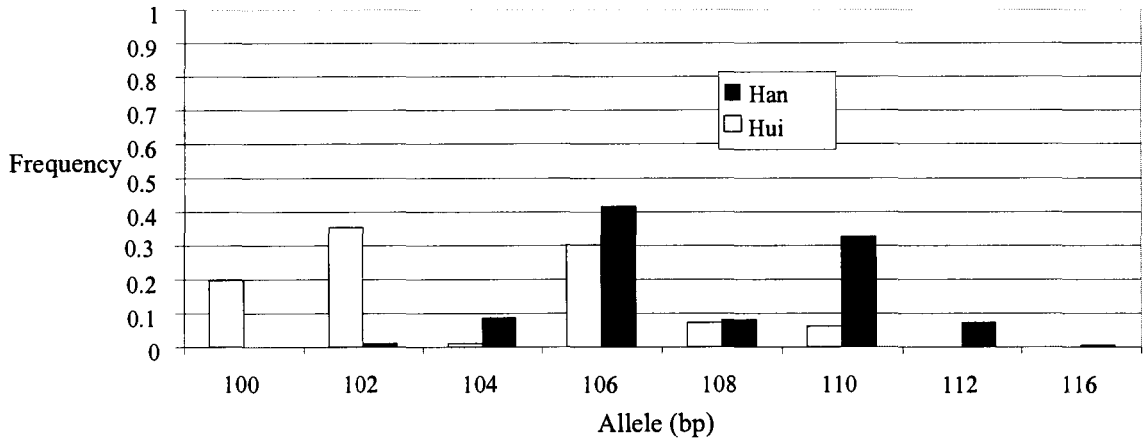
Phylip - Phylogeny program
Home page: http://evolution.genetics.washington.edu/phylip.html

Appendix B

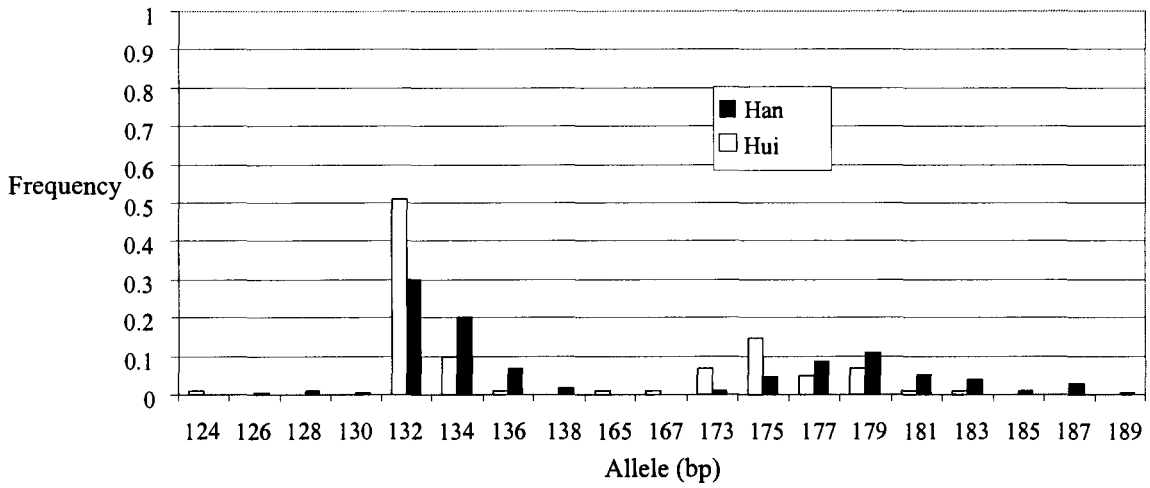
Autosomal allele distributions

B.1 Autosomal Frequency Distributions

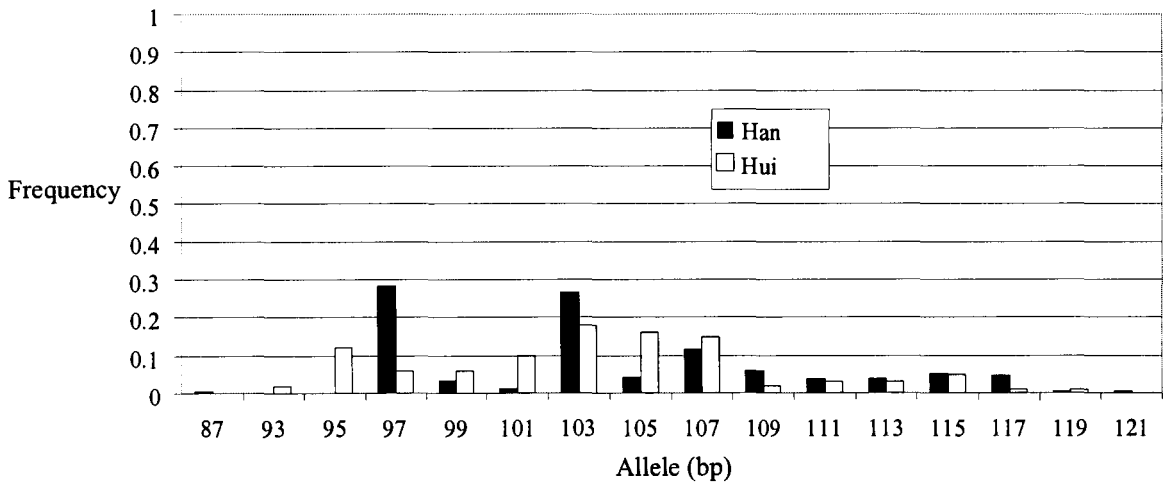
D13S126



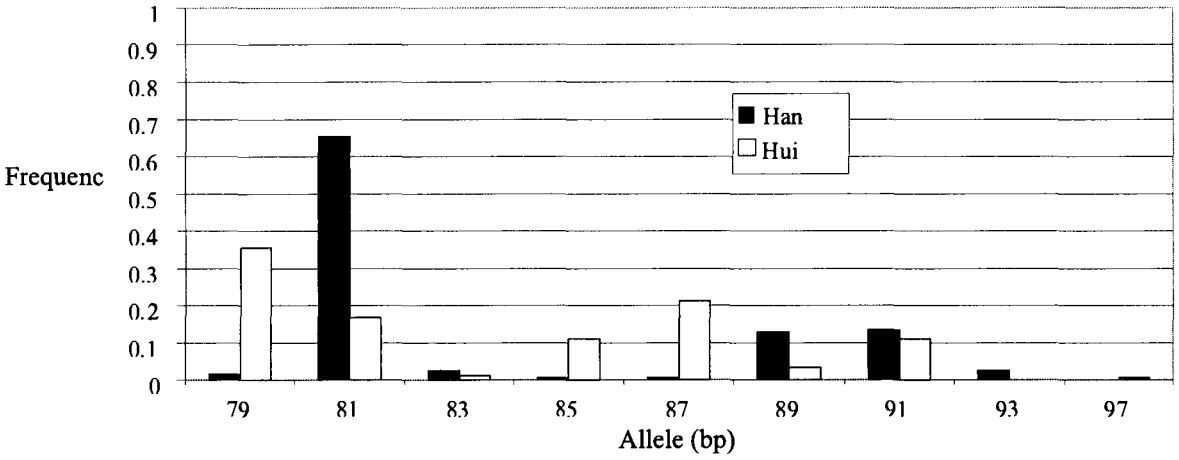
D13S133



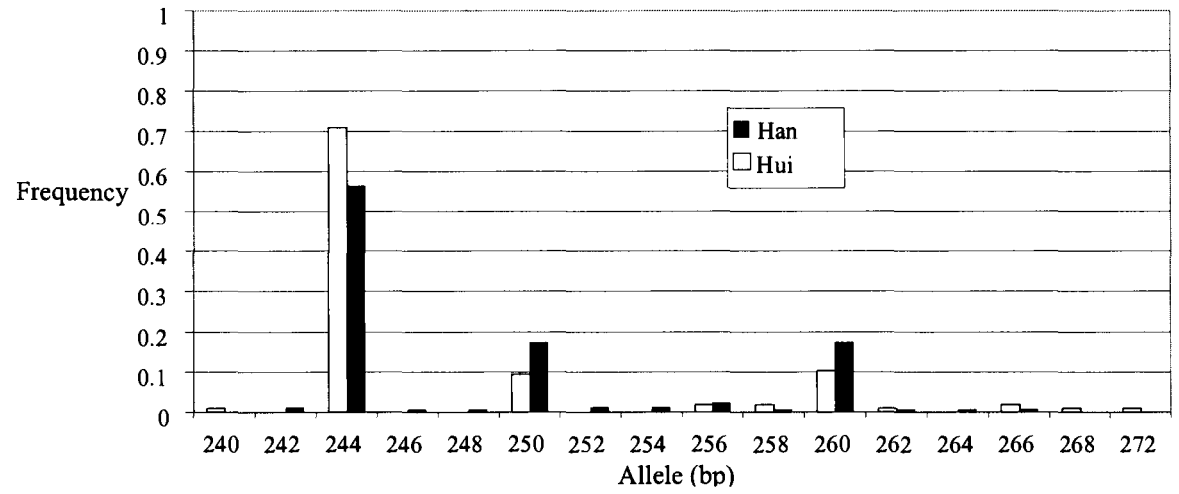
D13S192



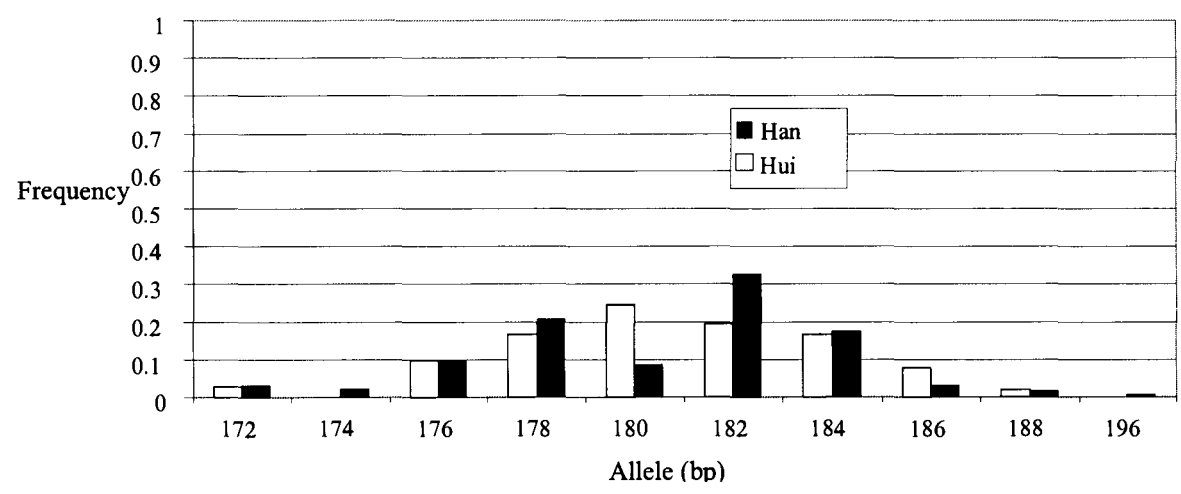
D13S270

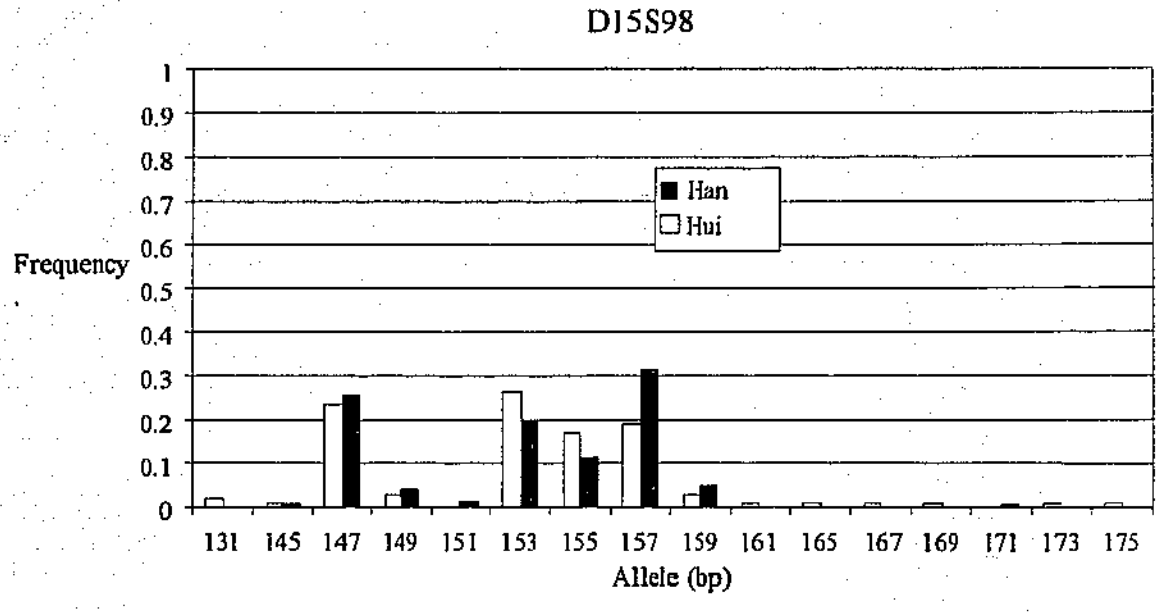


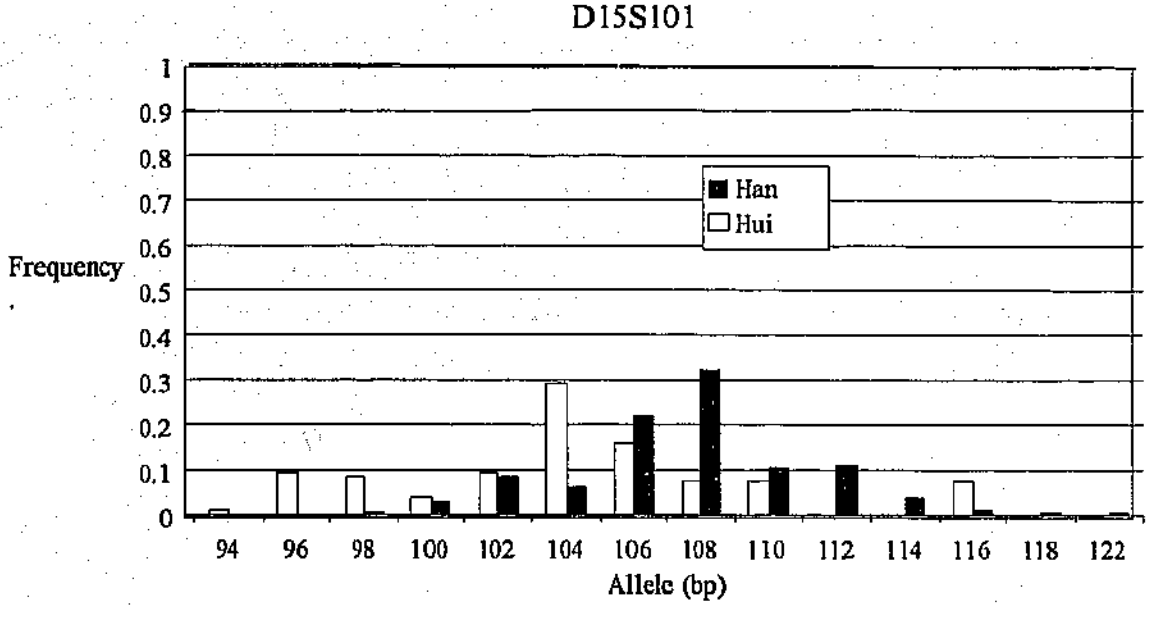
D15S11



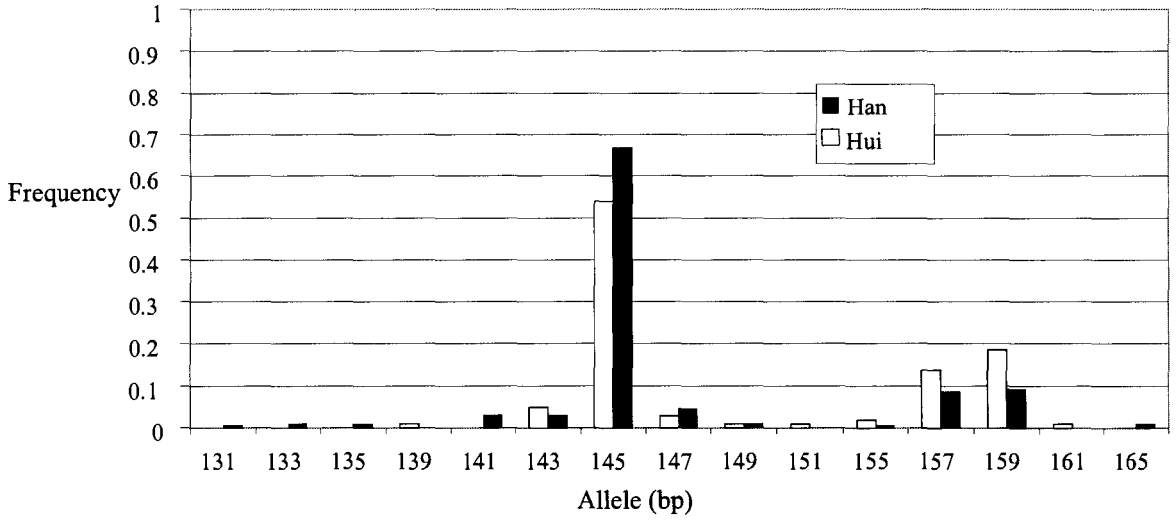
D15S97



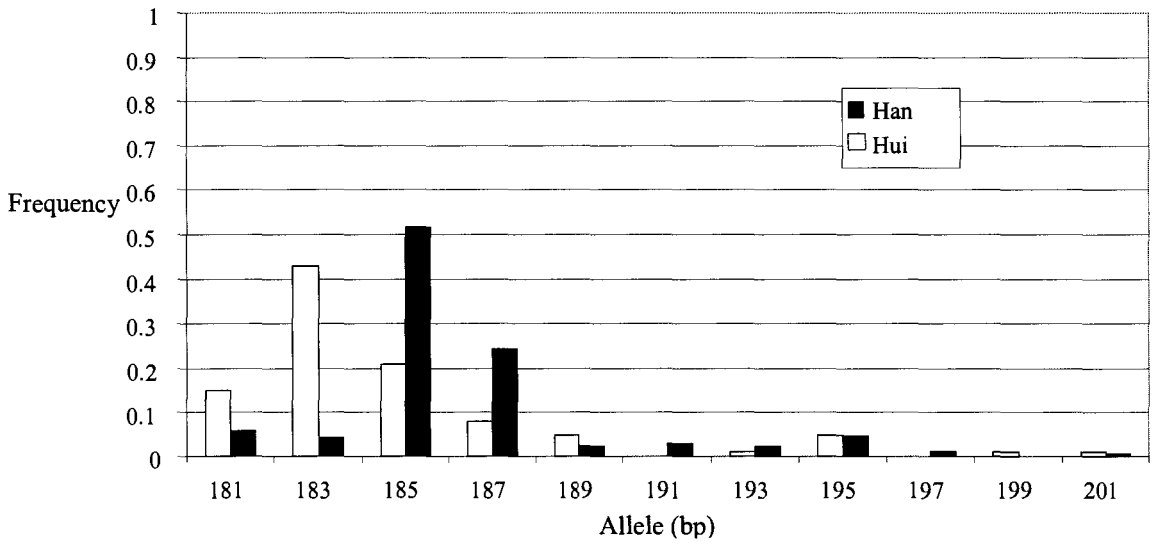

D15S98
Frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Han
Hui
131 145 147 149 151 153 155 157 159 161 165 167 169 171 173 175
Allele (bp)


D15S101
Frequency
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Han
Hui
94 96 98 100 102 104 106 108 110 112 114 116 118 122
Allele (bp)

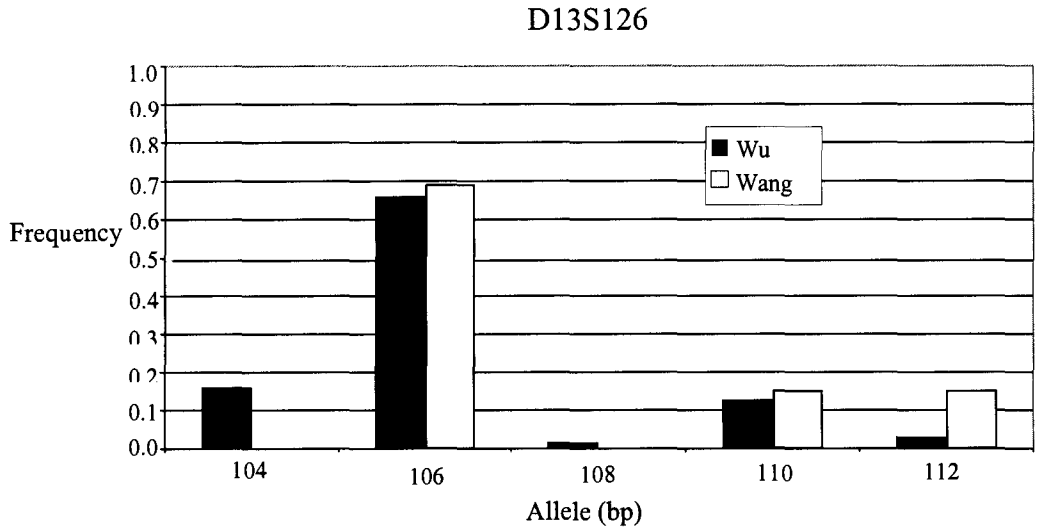
D15S108

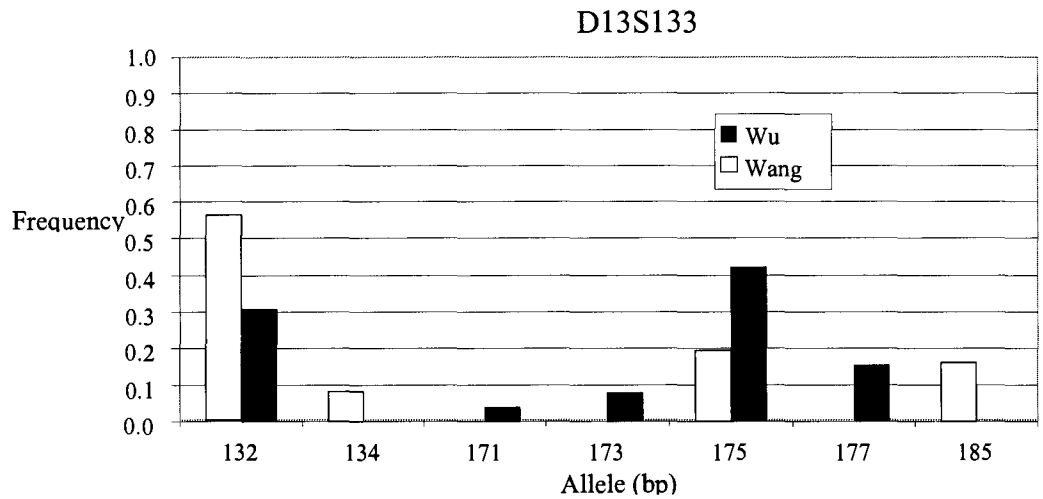


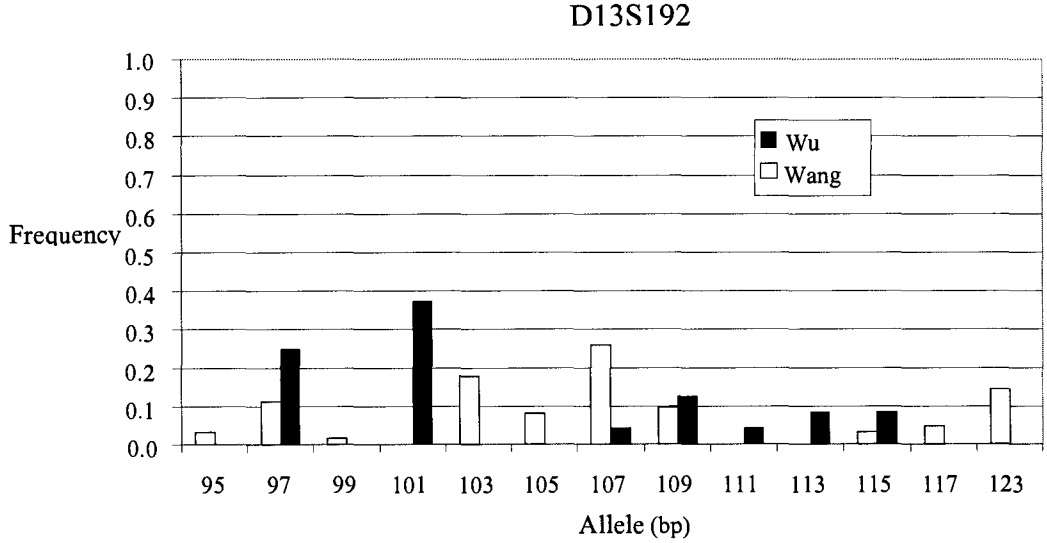
GABRB3



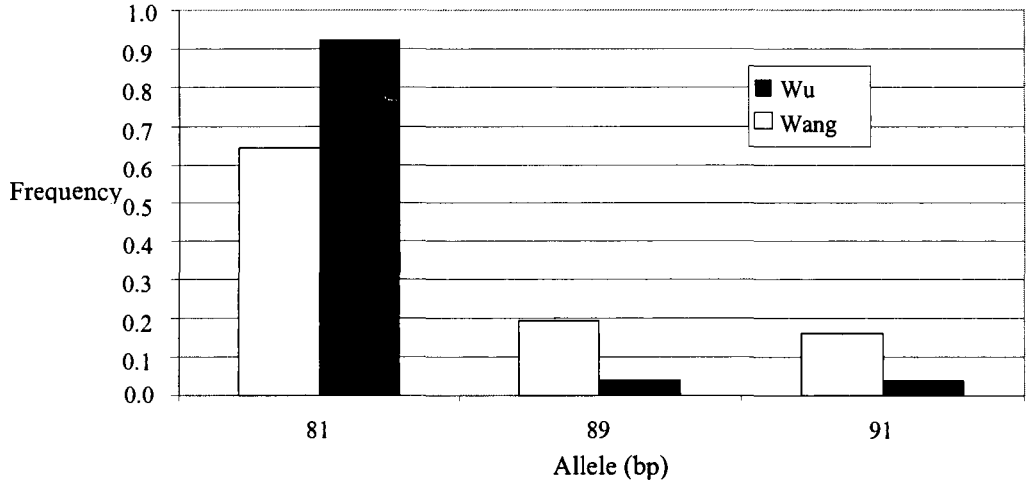
B.2 Pedigree autosomal allele distributions



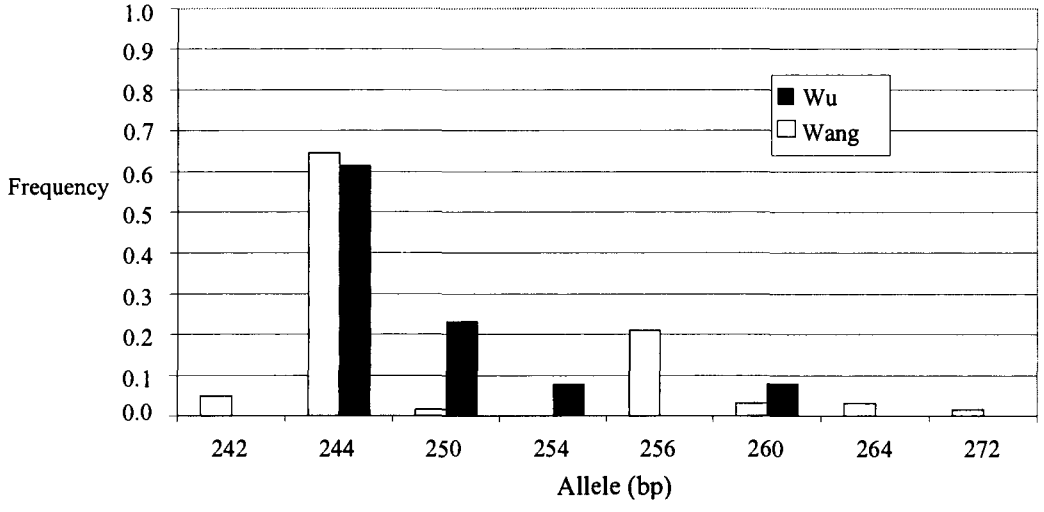




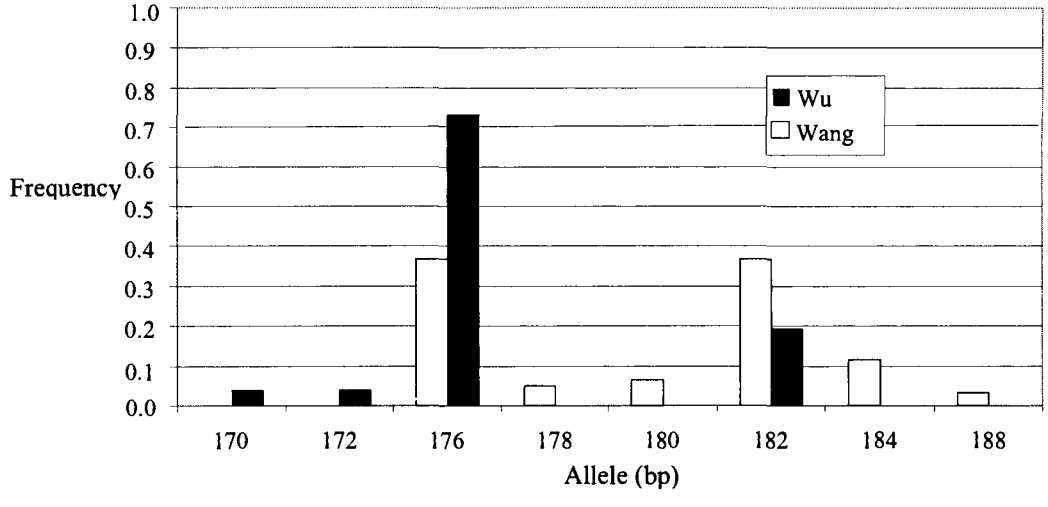
D13S270

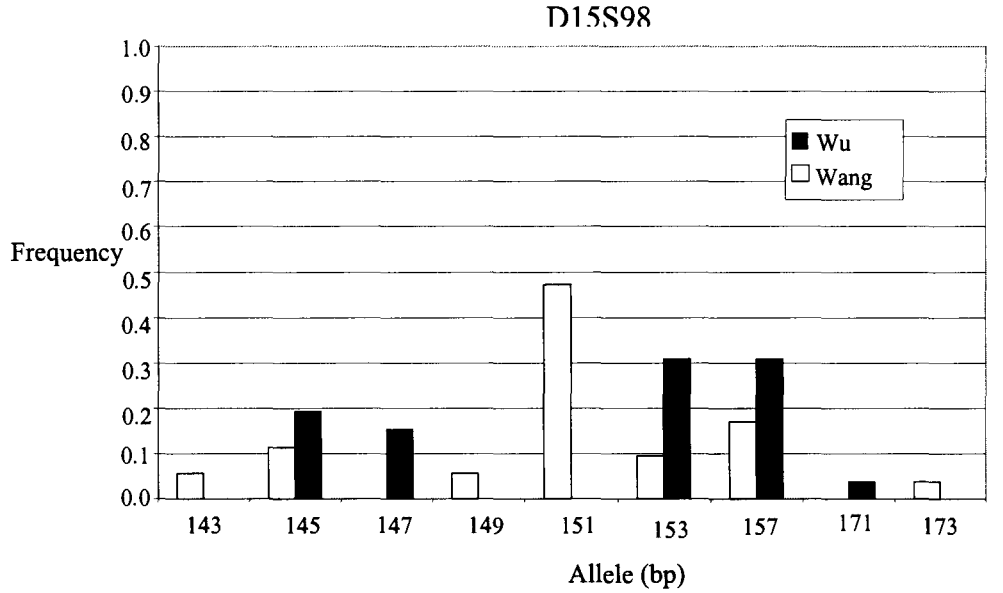


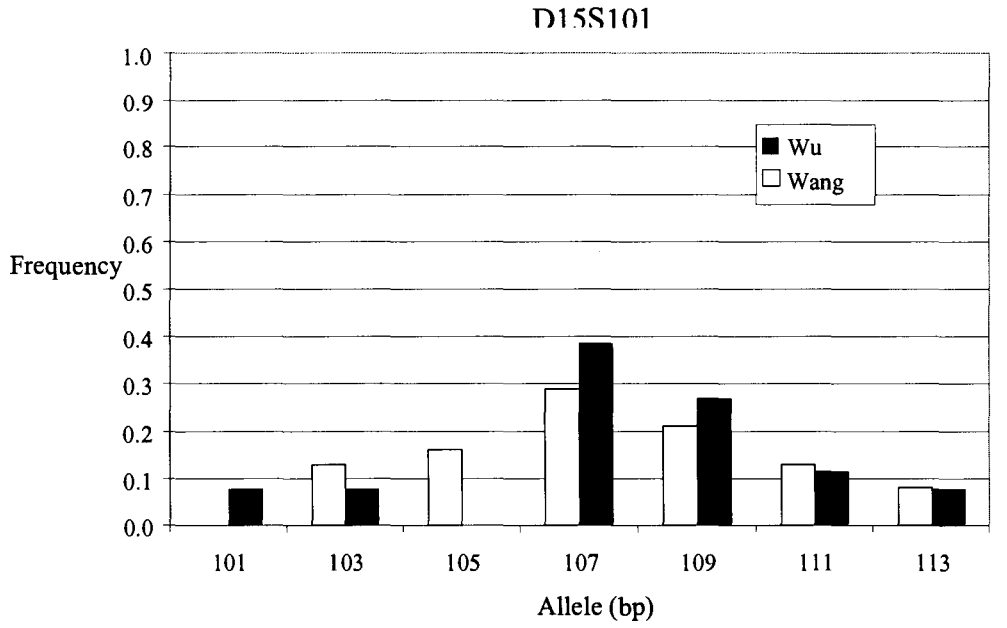
D15S11



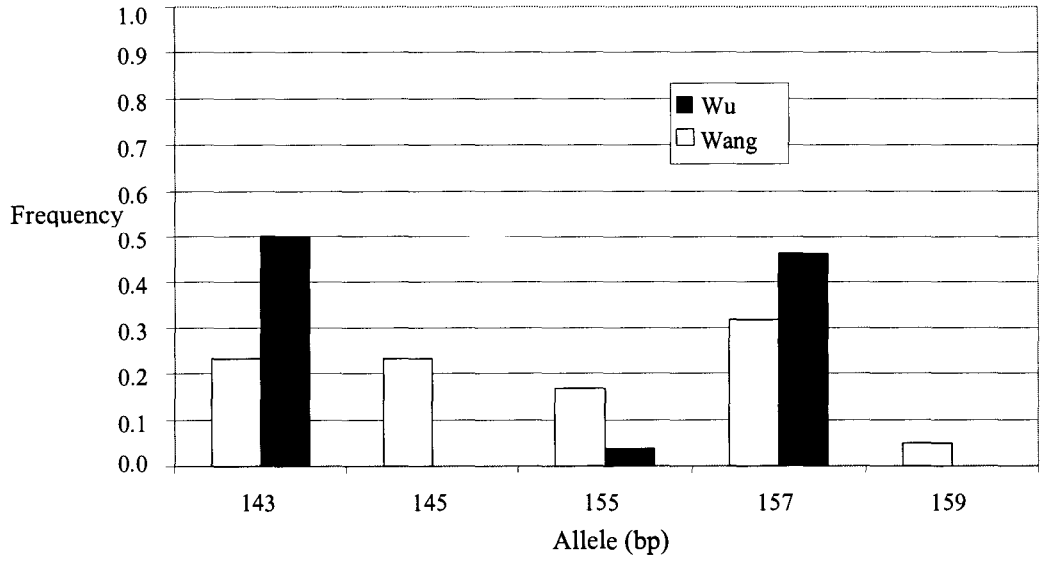
D15S97



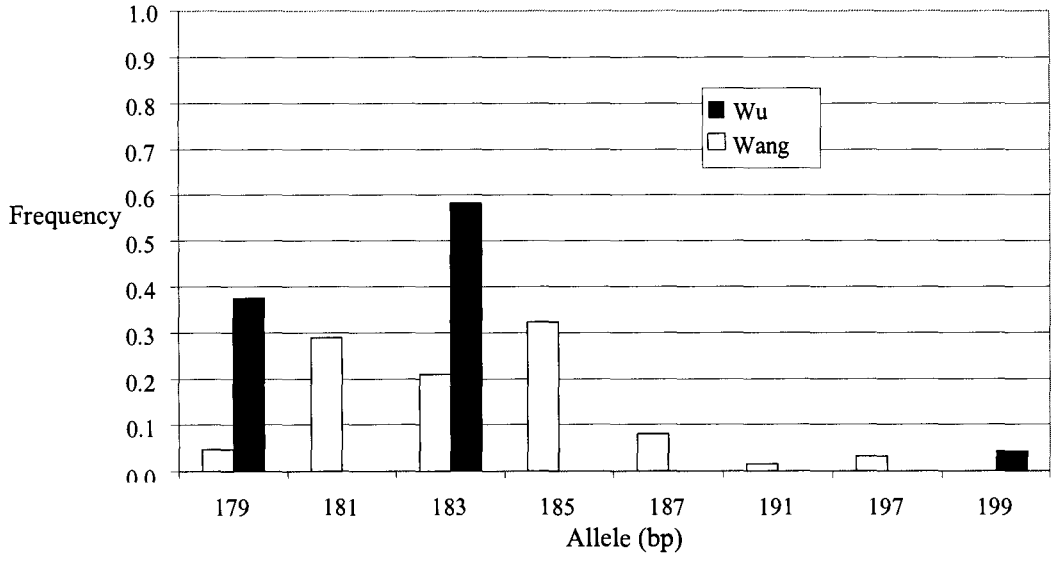

D15S98
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Wu
Wang
143
145
147
149
151
153
157
171
173
Allele (bp)


D15S101
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Wu
Wang
101
103
105
107
109
111
113
Allele (bp)

D15S108



GABRB3



Appendix C

Y-chromosome haplotypes and allele distributions

C.1. Random sample population haplotypes

(a) Han haplotypes

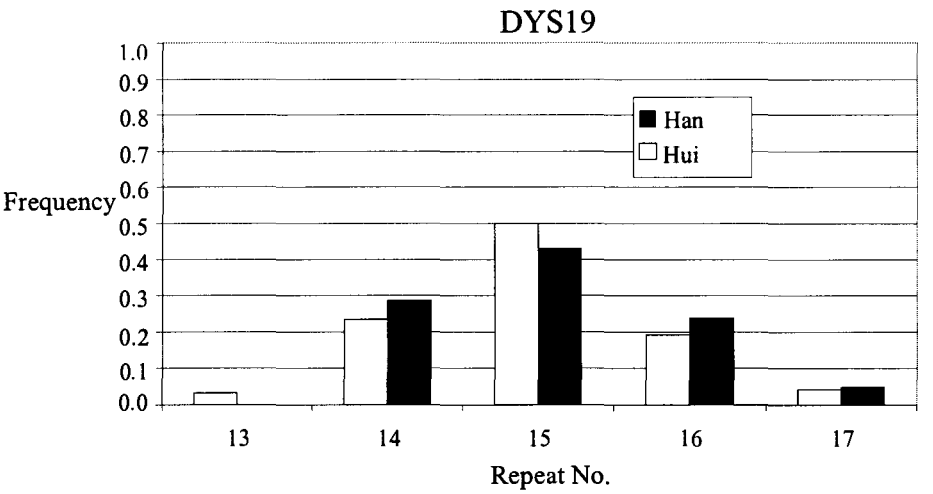
Haplotype	DYS19	DYS388	DYS389 I	DYS389 II	DYS390	DYS392	DYS393
1	14	10	12	28	25	16	12
2	14	13	13	29	24	16	14
3	16	12	12	31	21	15	13
4	15	14	12	29	22	14	15
5	15	10	12	29	22	14	12
6	15	12	12	31	24	15	12
7	15	12	12	28	24	15	12
8	15	12	12	27	24	14	13
9	15	12	12	28	26	14	12
10	15	13	12	27	26	14	12
11	15	12	12	29	26	14	14
12	?	12	12	29	?	15	14
13	14	11	13	28	22	15	14
14	14	10	13	31	25	16	12
15	15	10	13	29	?	16	12
16	15	12	12	29	24	16	13
17	15	12	12	28	25	16	13
18	15	10	12	30	25	16	12
19	14	?	13	29	25	16	12
20	14	12	14	31	24	14	12
21	?	?	12	28	26	14	12
22	15	13	14	30	25	13	13
23	17	12	12	28	25	15	13
24	17	12	12	28	24	15	14
25	15	12	12	28	24	15	14
26	14	12	14	31	24	14	13
27	15	10	13	30	25	14	12
28	15	12	13	30	24	13	12
29	15	13	13	30	25	14	12
30	15	12	12	28	25	14	12
31	16	12	12	27	26	14	12
32	16	12	13	27	24	14	14
33	14	12	12	28	24	14	12
34	14	13	12	28	24	13	12
35	15	10	?	?	25	14	12
36	15	12	13	29	26	14	12
37	13	15	14	30	24	15	14
38	13	12	13	30	25	14	14
39	16	17	14	30	25	14	12
40	14	13	12	28	25	15	12
41	14	14	13	28	26	16	13
42	14	10	12	28	24	15	12
43	14	12	13	29	?	12	12
44	15	10	14	30	24	12	15
45	15	12	14	30	25	12	14
46	14	14	12	28	24	15	12
47	14	13	12	28	24	15	14

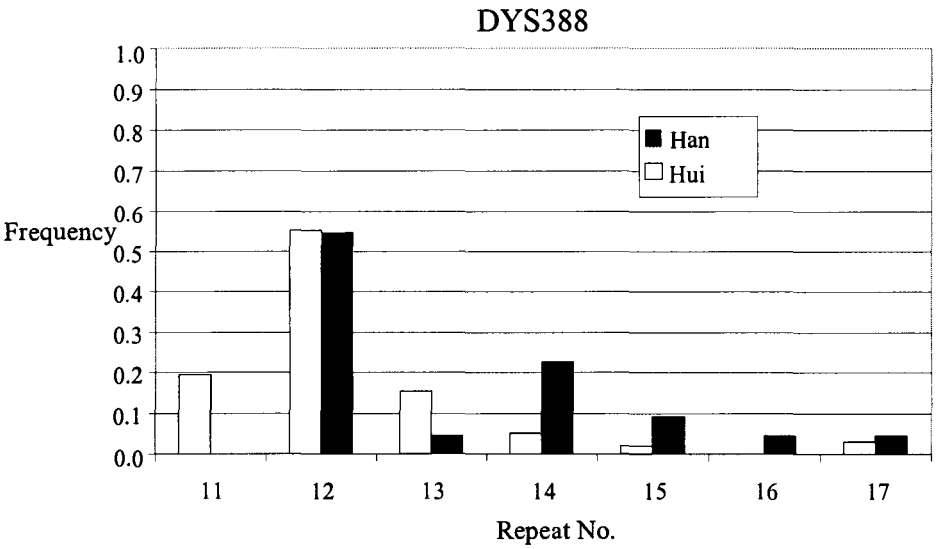
48	17	10	12	28	25	14	12
49	15	10	12	28	?	14	12
50	15	12	14	32	24	14	14
51	15	12	14	32	24	?	15
52	15	15	14	32	24	14	15
53	15	12	12	27	25	14	13
54	14	12	12	28	24	15	12
55	14	12	13	29	24	12	13
56	14	10	14	30	25	12	14
57	?	13	13	?	?	12	15
58	16	17	12	29	26	13	12
59	14	12	11	31	26	15	12
60	14	13	13	30	26	13	12
61	14	12	13	?	?	14	12
62	?	?	12	28	22	14	12
63	16	12	12	28	22	14	12
64	16	12	12	?	25	13	13
65	16	12	13	28	26	14	12
66	16	12	14	32	26	14	13
67	16	12	12	29	22	15	12
68	16	12	13	29	26	14	13
69	?	12	12	28	26	14	12
70	?	14	12	28	26	14	12
71	?	12	13	?	26	14	12
72	16	12	15	?	25	14	12
73	15	12	15	29	24	14	12
74	16	12	12	28	?	14	12
75	16	17	12	28	24	13	12
76	?	12	13	31	23	12	13
77	16	14	13	28	23	14	12
78	15	?	12	28	?	15	13
79	14	12	12	27	24	14	12
80	15	10	?	28	25	14	12
81	15	13	13	29	25	15	13
82	15	10	14	31	25	15	12
83	15	13	14	31	23	12	13
84	15	12	12	29	25	13	13
85	13	12	12	28	22	13	12
86	16	12	12	28	?	14	12
87	16	12	12	30	23	14	12
88	14	12	13	29	24	15	13
89	14	10	12	28	22	15	12
90	14	12	12	28	22	14	12
91	17	11	13	28	20	15	12
92	15	13	12	28	21	12	13
93	15	12	?	?	?	12	13
94	15	12	14	?	?	12	13
95	16	12	14	33	21	12	13
96	15	13	14	30	22	12	13
97	15	12	12	29	26	12	14
98	15	13	13	29	21	?	12
99	15	12	13	29	23	?	12
100	15	10	12	28	21	?	12
101	15	12	13	?	?	?	13
102	15	13	14	30	?	?	13

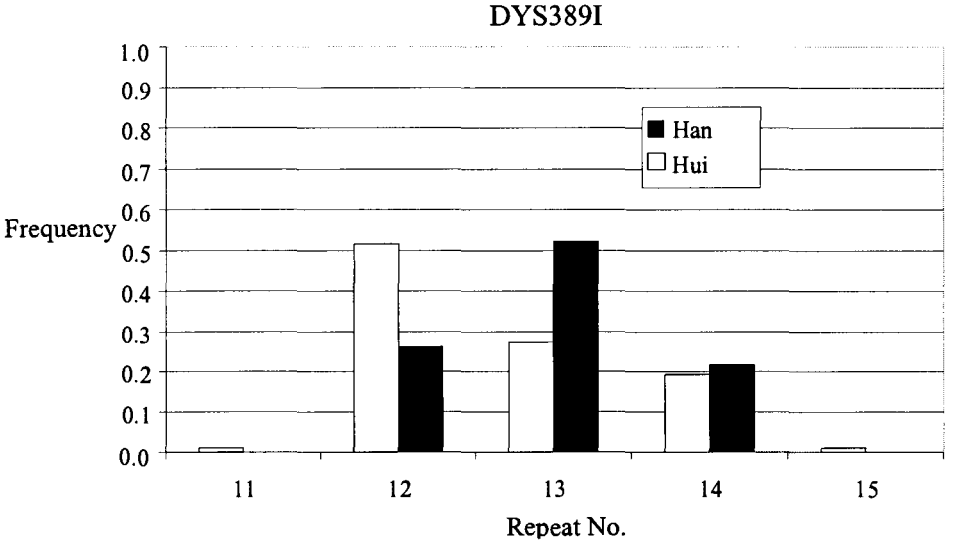
(b) Hui haplotypes

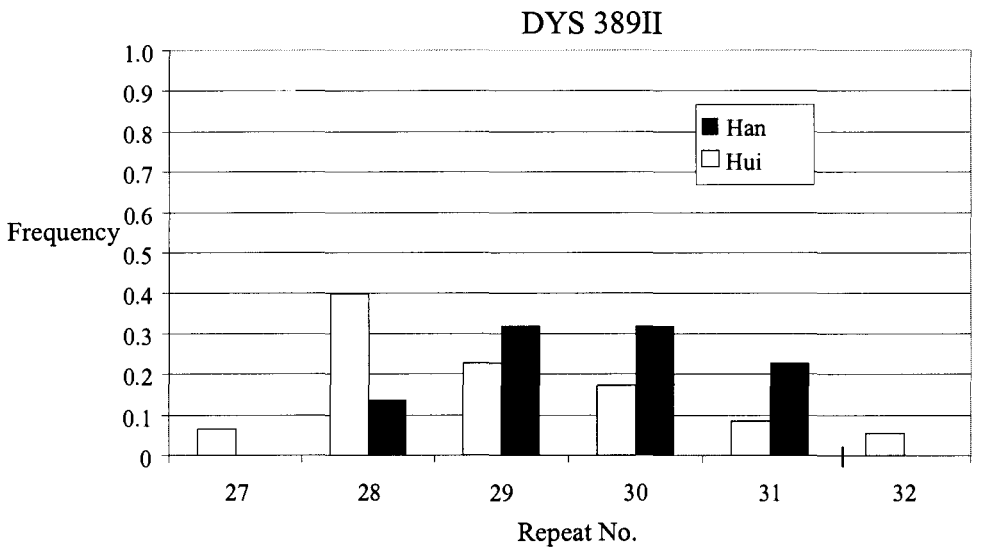
Haplotypes	DYS19	DYS388	DYS389 I	DYS389 II	DYS390	DYS392	DYS393
Hui 1	16	12	12	28	24	14	12
Hui 2	15	0	12	28	24	?	12
Hui 3	16	12	12	?	?	?	?
Hui 4	15	12	13	31	25	?	13
Hui 5	16	12	13	31	24	14	13
Hui 6	14	12	?	?	24	14	?
Hui 7	14	12	13	31	23	14	13
Hui 8	15	16	13	29	23	14	12
Hui 9	15	12	13	31	24	12	13
Hui 10	16	12	13	29	23	12	13
Hui 11	14	17	13	30	23	12	12
Hui 12	14	12	13	31	23	14	13
Hui 13	15	14	13	29	24	12	15
Hui 14	14	13	14	31	23	12	12
Hui 15	15	14	13	29	24	12	15
Hui 16	14	15	13	30	26	14	13
Hui 17	16	12	14	30	25	14	13
Hui 18	17	14	12	29	24	14	12
Hui 19	15	14	14	30	23	12	14
Hui 20	15	12	14	30	24	12	14
Hui 22	15	12	12	28	22	15	13
Hui 23	14	15	12	29	21	12	15
Hui 24	?	13	13	29	23	12	12
Hui 25	16	12	14	31	25	14	?

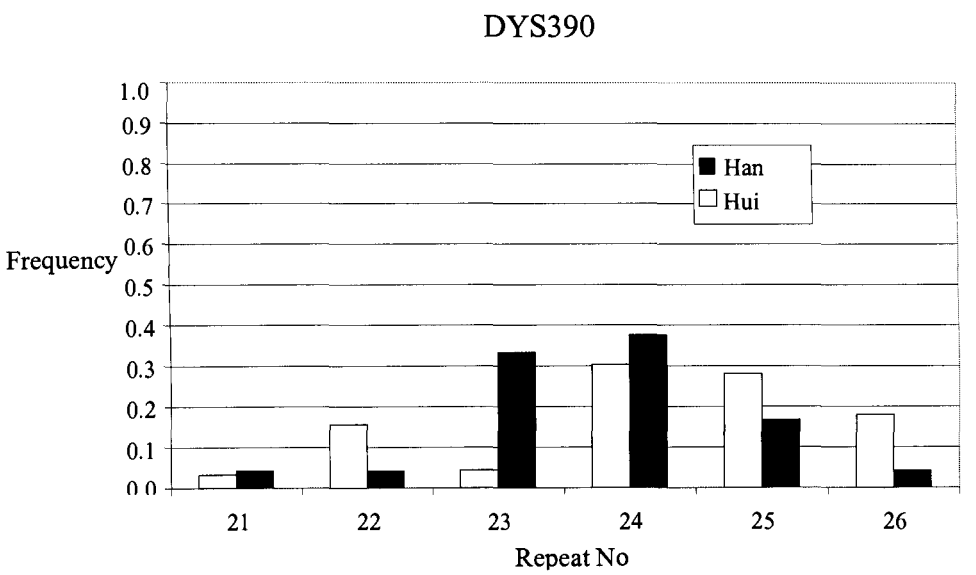
C.2. Allele frequency distributions of random sample populations

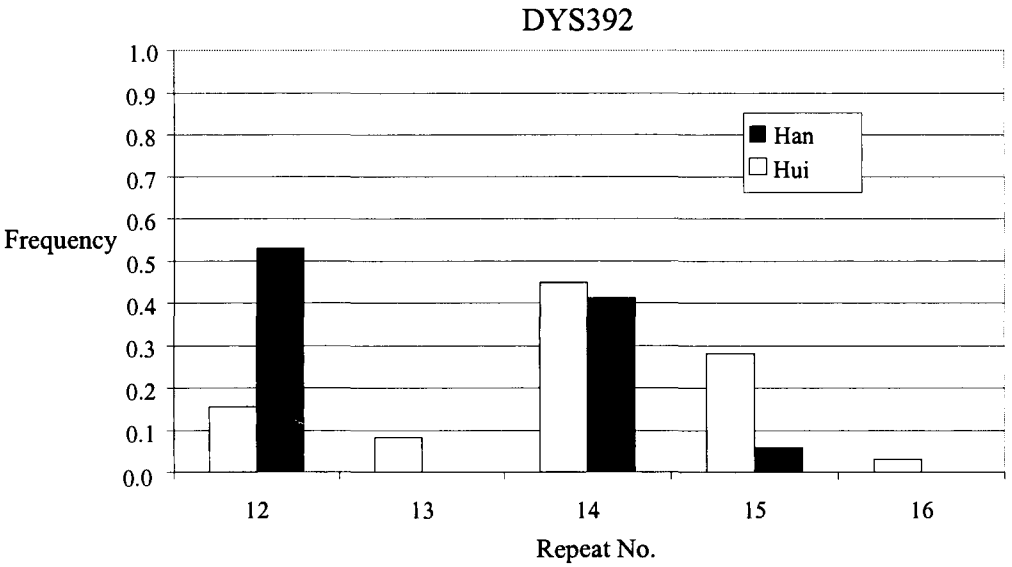


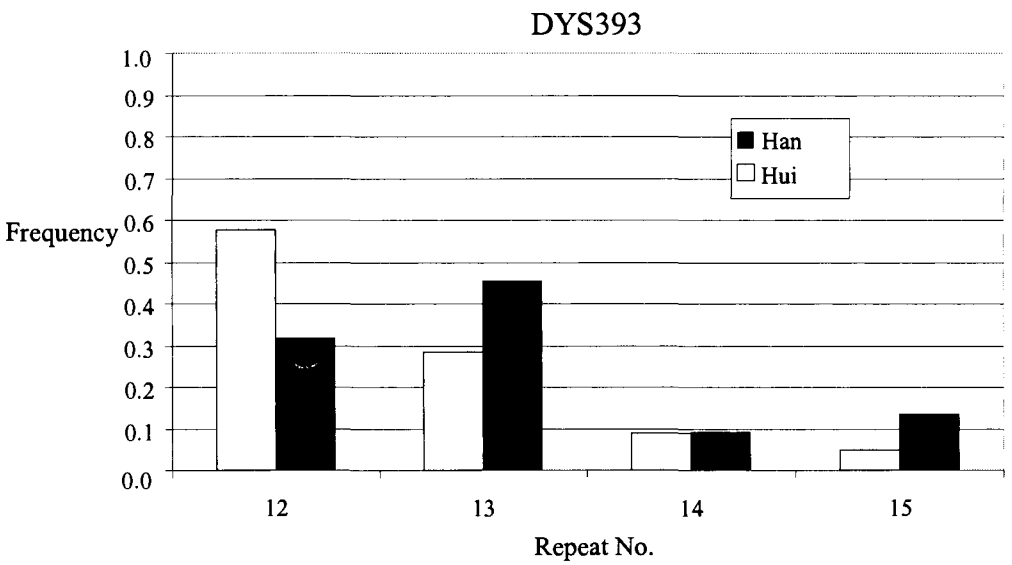





DYS 389II
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Frequency
Han
Hui
27
28
29
30
31
32
Repeat No.


DYS390
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0 0
Frequency
Han
Hui
21
22
23
24
25
26
Repeat No


DYS392
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Han
Hui
12
13
14
15
16
Repeat No.


DYS393
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Han
Hui
12
13
14
15
Repeat No.

C.3 Pedigree haplotypes

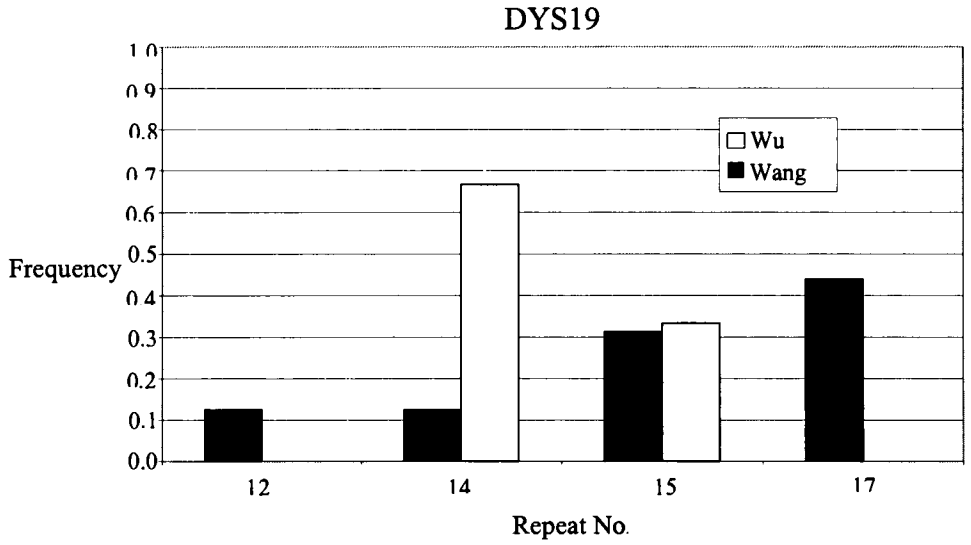
(a) Wang haplotypes

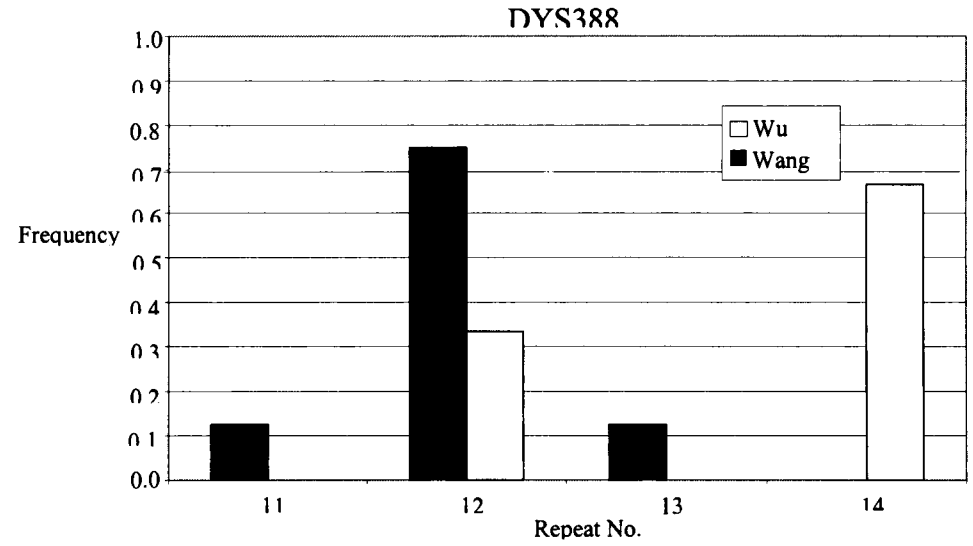
Sample No	DYS19	DYS388	DYS389I	DYS389II	DYS390	DYS392	DYS393
Wang 1	17	12	12	27	23	14	11
Wang 3	17	11	12	30	24	14	11
Wang 6	17	11	12	30	24	14	11
Wang 7	12	14	14	30	25	14	12
Wang 8	12	14	14	30	25	14	12
Wang 9	17	12	12	27	23	14	11
Wang 11	15	12	13	31	24	11	11
Wang 12	15	12	13	31	24	11	11
Wang 14	17	12	12	27	23	14	11
Wang 16	17	12	12	27	23	14	11
Wang 18	17	12	12	27	23	14	11
Wang 20	15	12	12	27	22	15	12
Wang 22	15	12	12	27	22	15	12
Wang 26	15	12	12	27	22	15	11
Wang 29	14	12	12	27	23	13	11
Wang 30	14	12	12	27	23	13	11

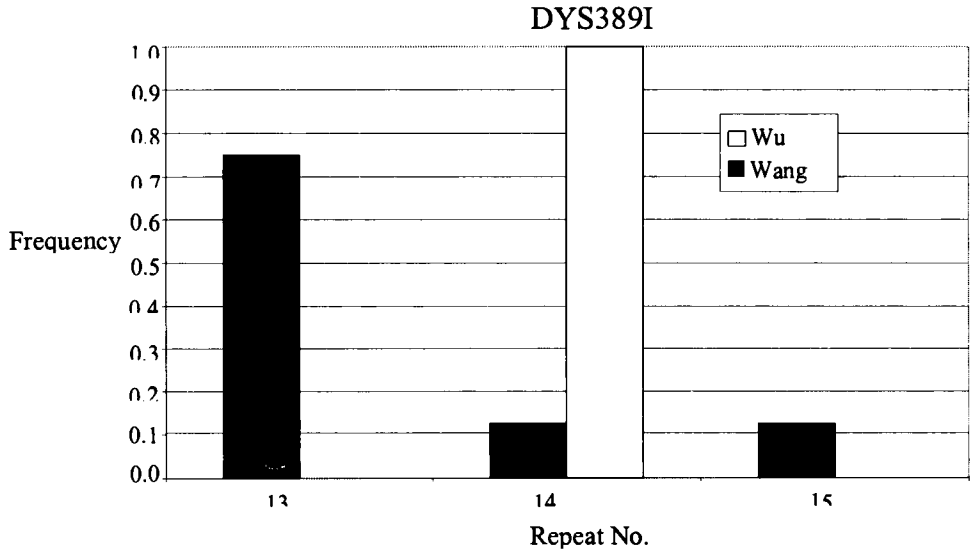
(b) Wu haplotypes

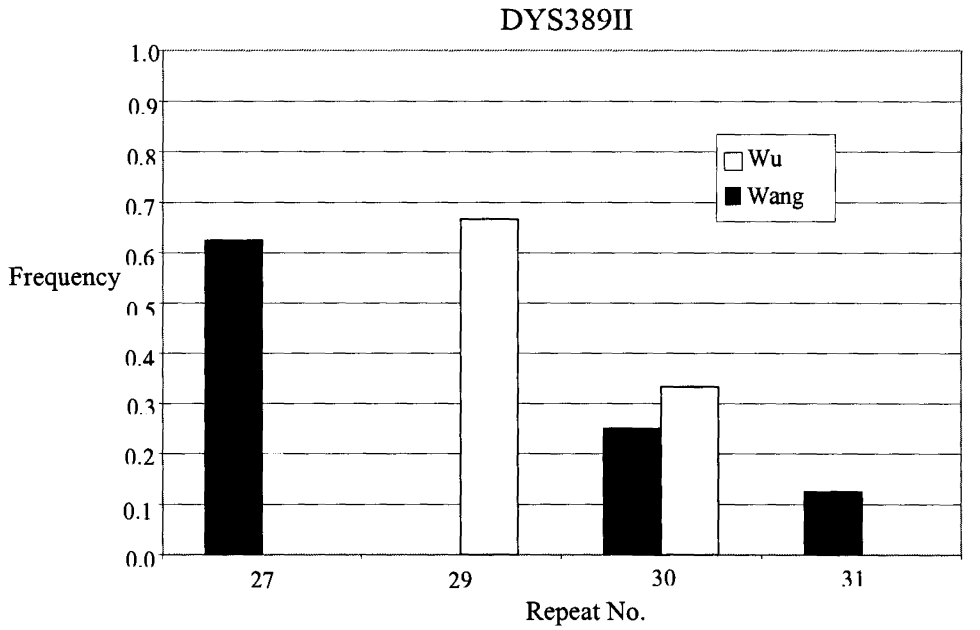
Sample No	DYS19	DYS388	DYS389I	DYS389II	DYS390	DYS392	DYS393
Wu 1	14	15	13	29	24	10	11
Wu 6	15	12	13	30	25	10	12
Wu 9	15	12	13	30	25	10	12
Wu 10	14	15	13	29	24	10	12
Wu 11	14	15	13	29	24	10	11
Wu 14	14	15	13	29	24	10	11

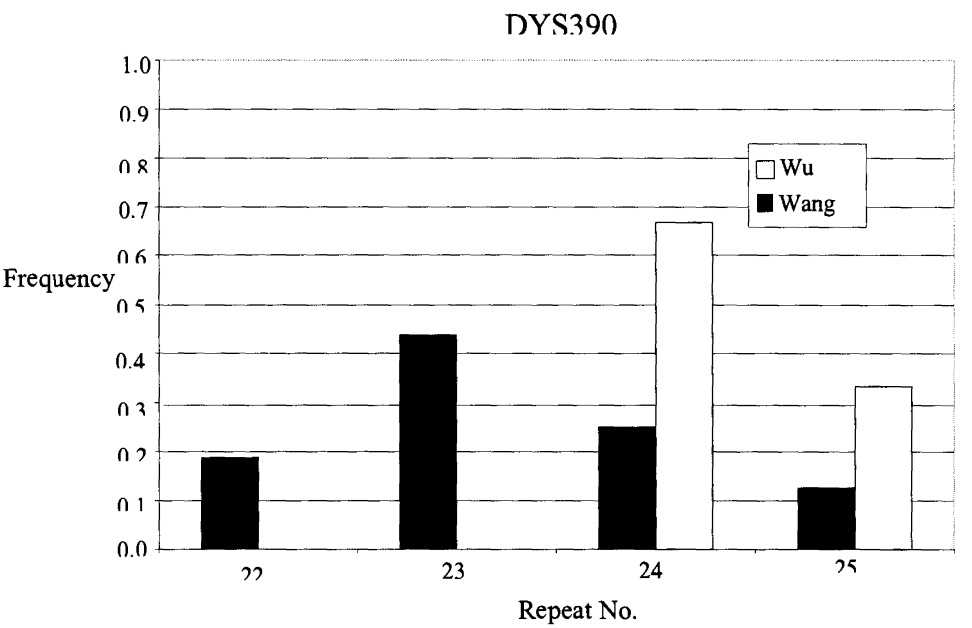
C.4 Allele frequency distributions of the pedigrees

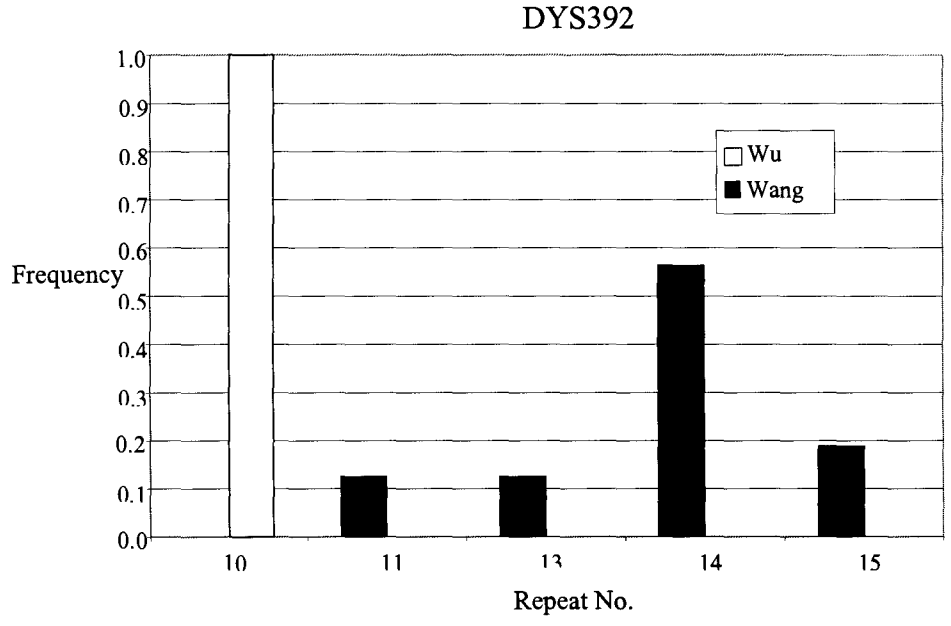


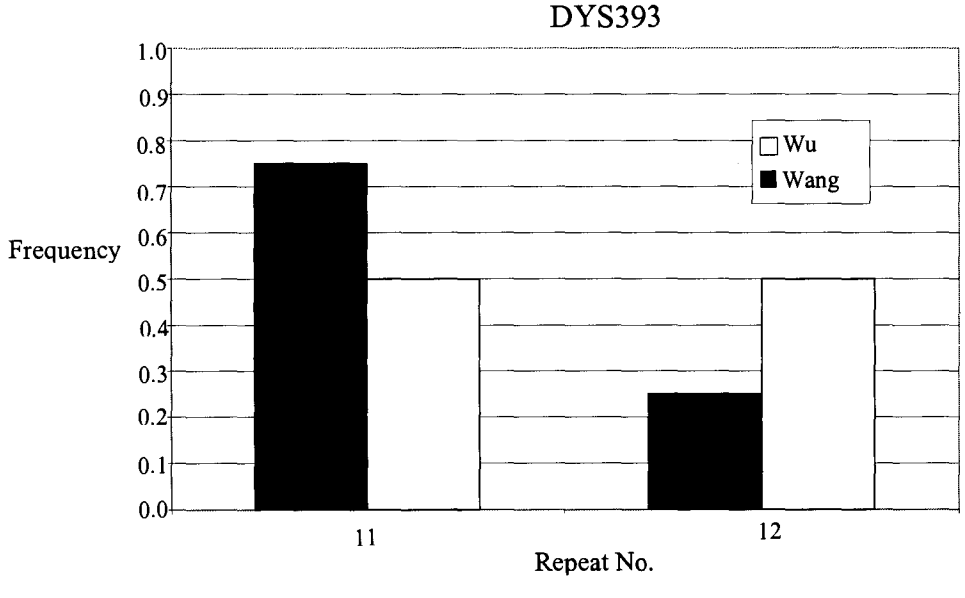





DYS389II
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Wu
Wang
27
29
30
31
Repeat No.


DYS390
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Wu
Wang
22
23
24
25
Repeat No.


DYS392
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Wu
Wang
10
11
13
14
15
Repeat No.


DYS393
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Frequency
Wu
Wang
11
12
Repeat No.

Appendix D

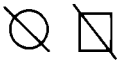
Wang and Wu pedigrees

Key:

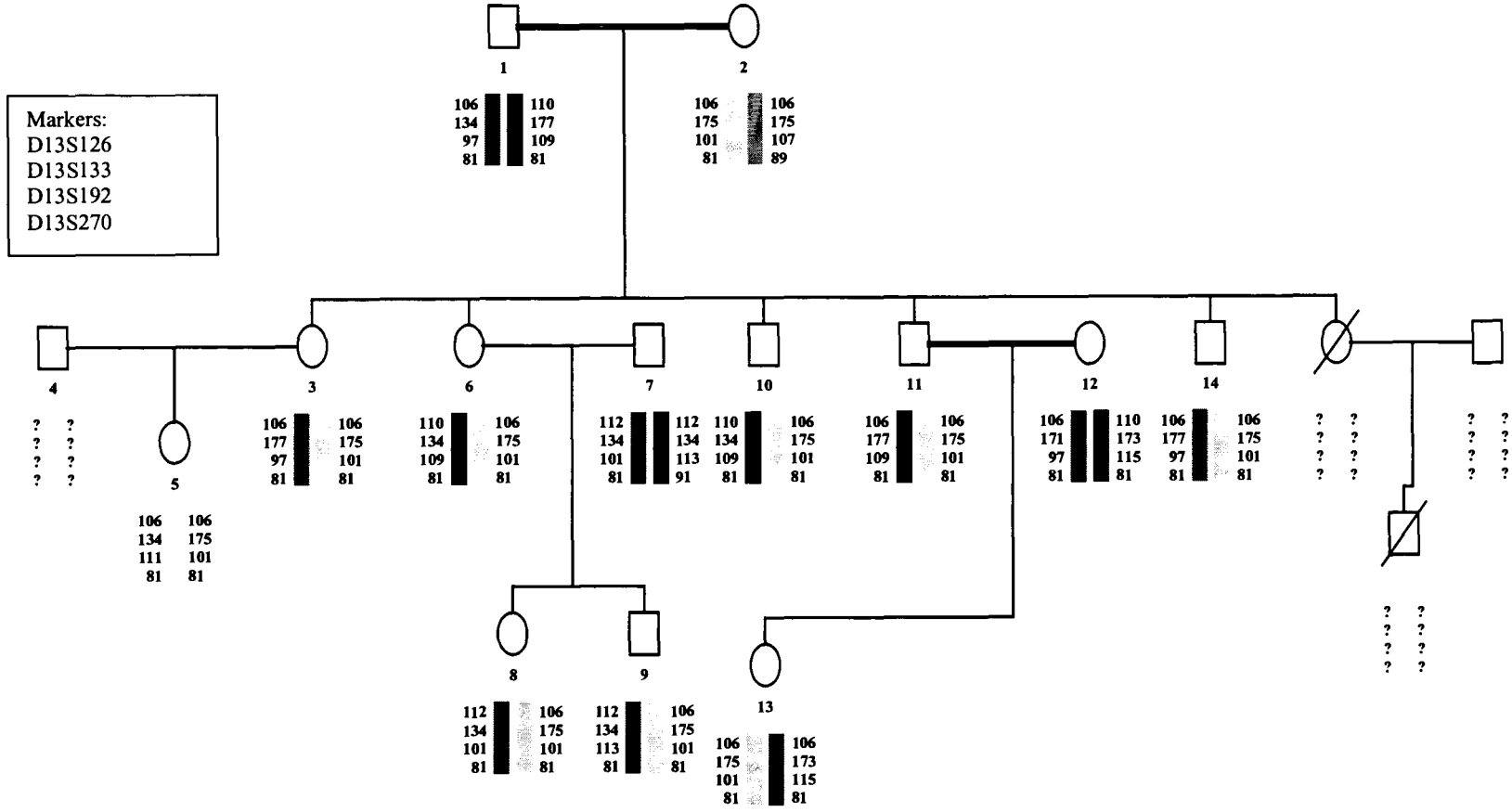
 Male individual

 Female individual

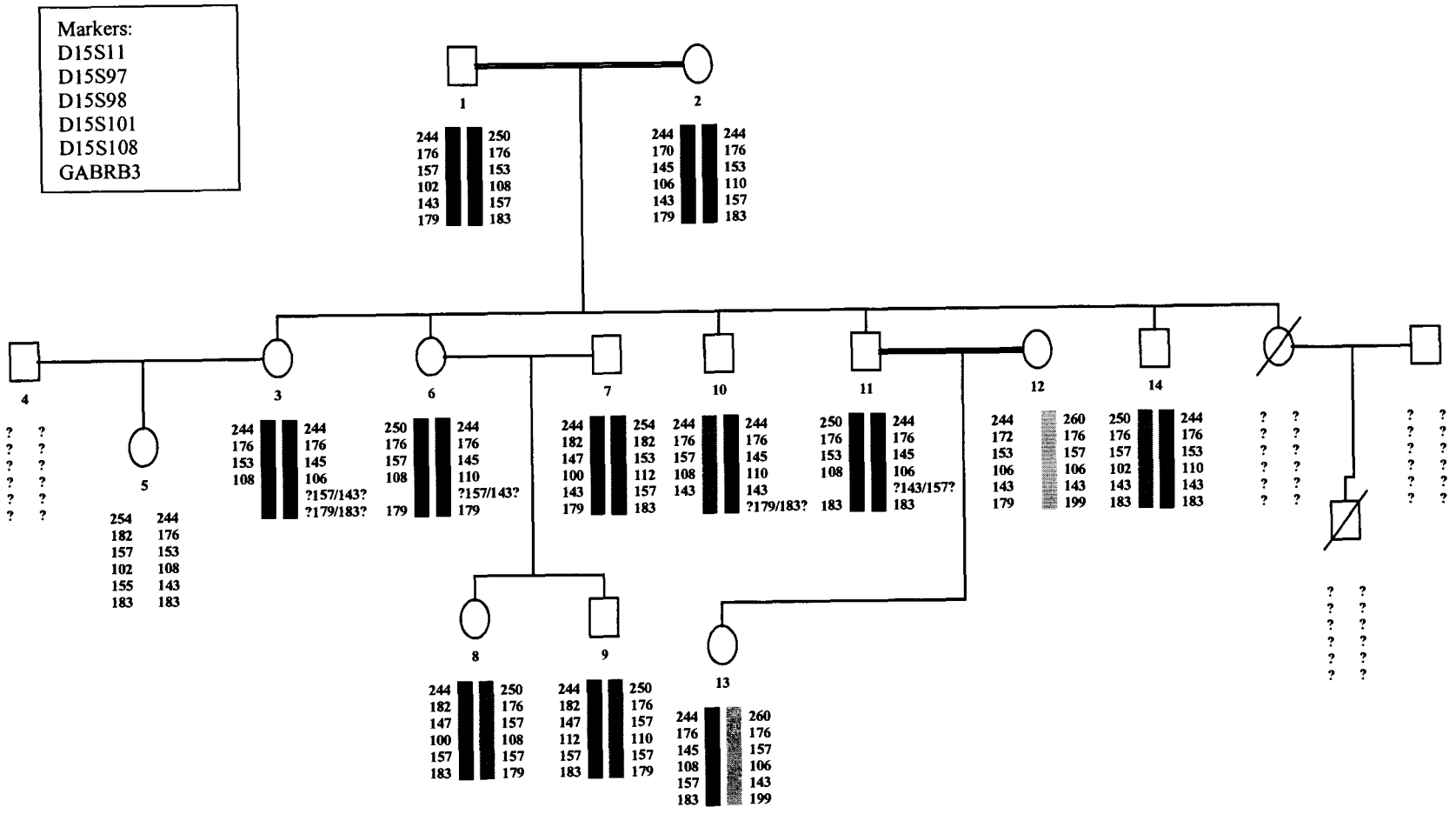
 Consanguineous marriage

 Deceased individual

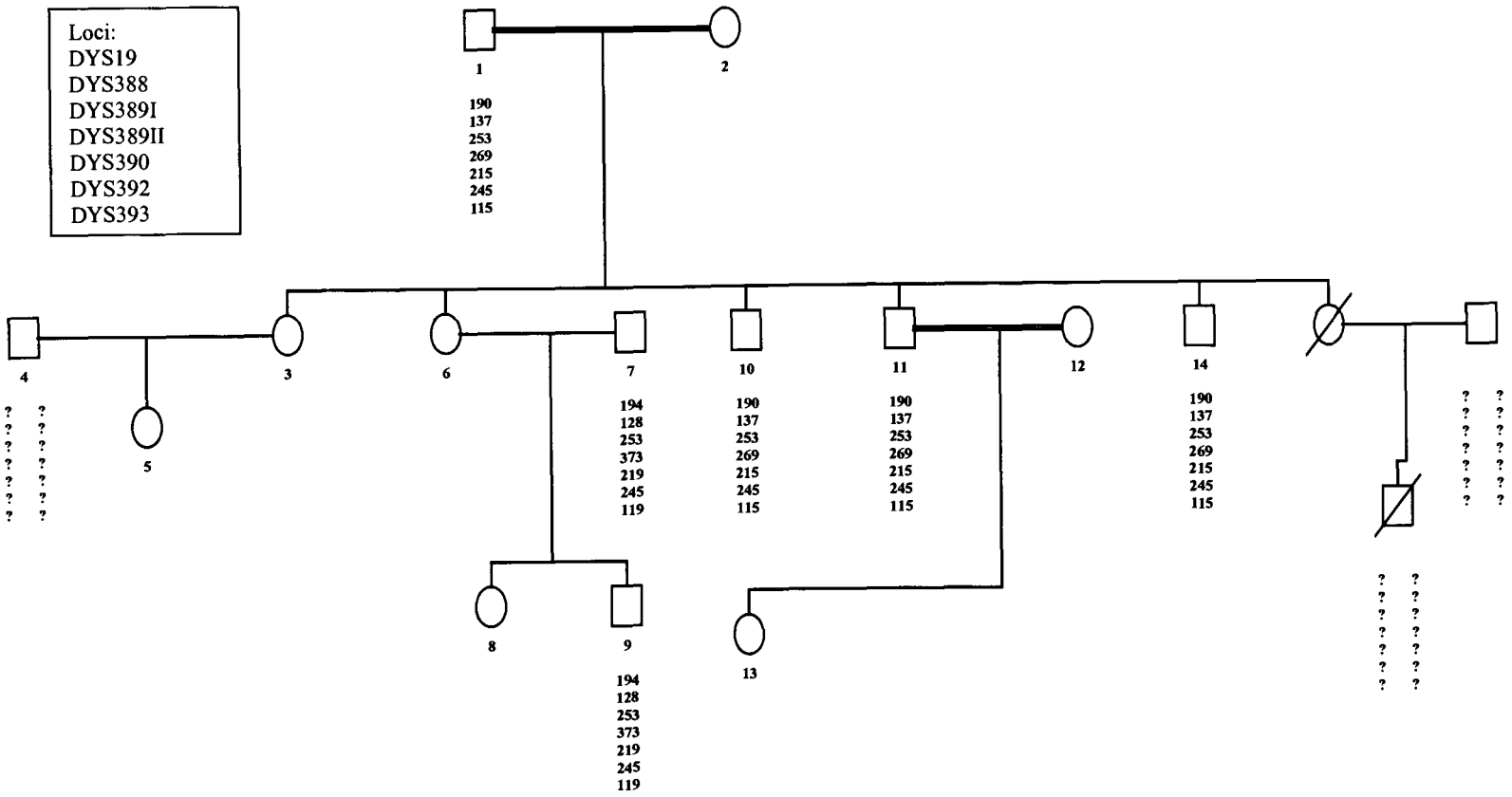
D.1 Wu pedigree – chromosome 13 genotypes



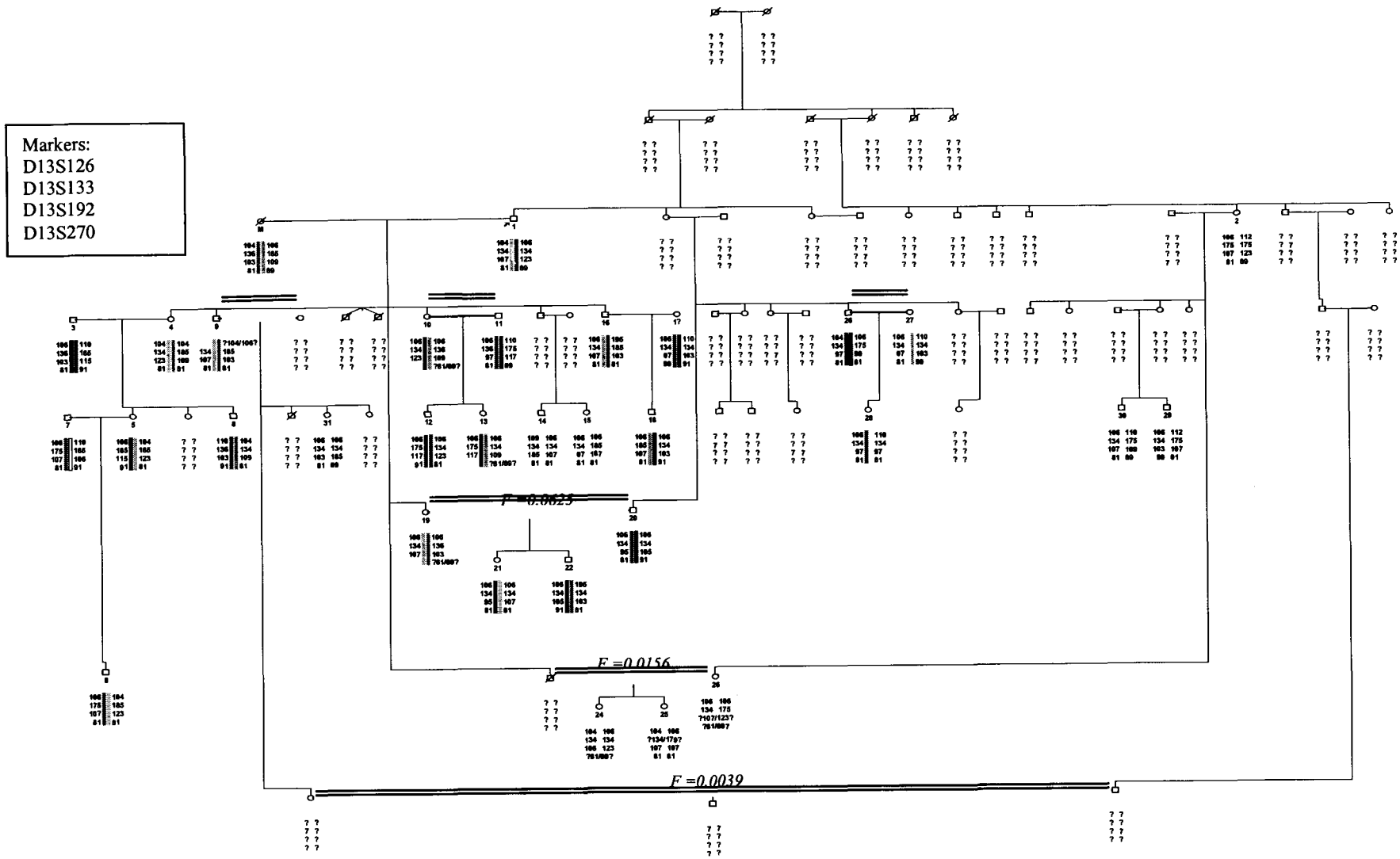
D.2 Wu pedigree – chromosome 15 genotypes



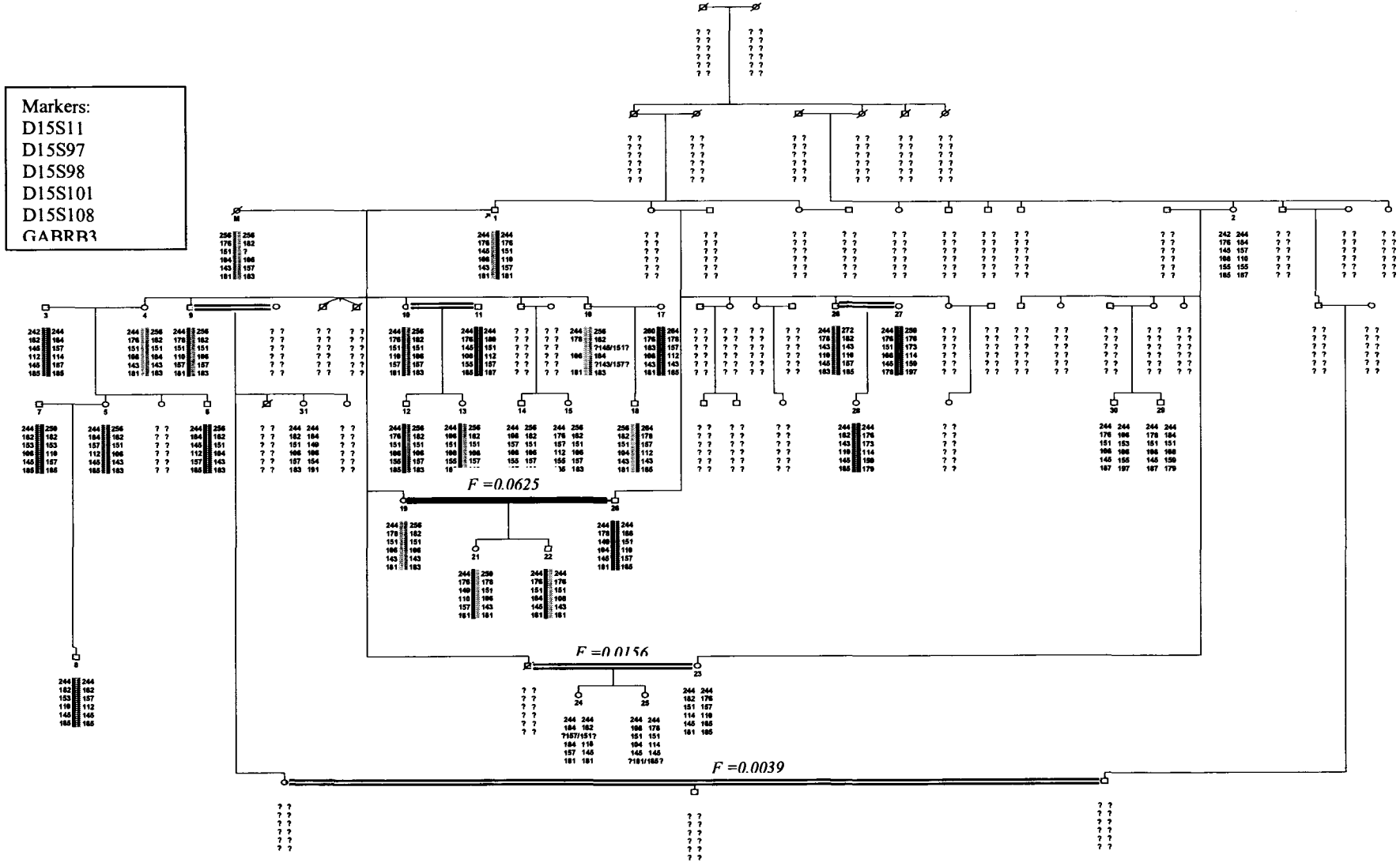
D.3 Wu pedigree – Y-chromosome genotypes



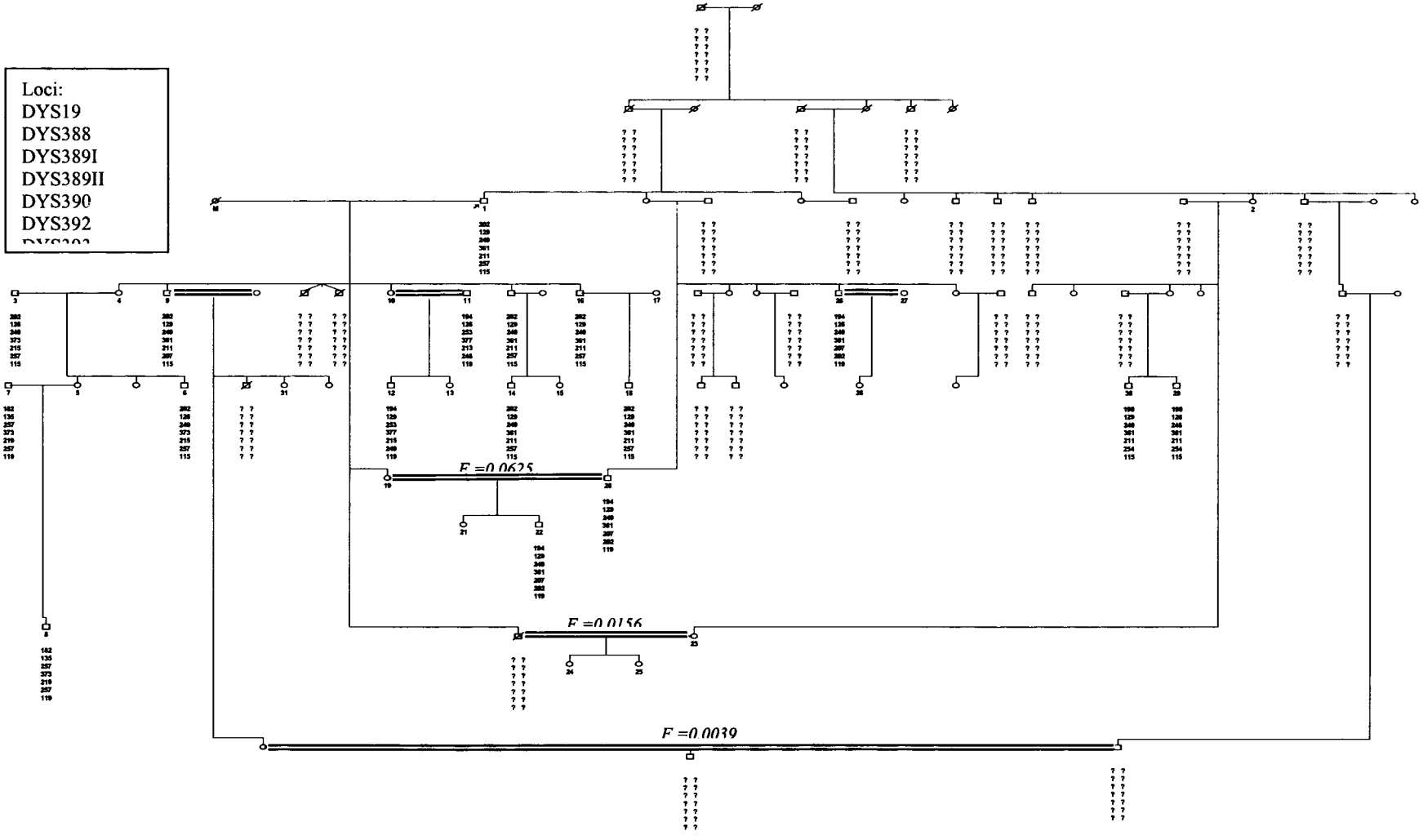
D.4 Wang pedigree – chromosome 13 genotypes



D.5 Wang pedigree – chromosome 15 genotypes



D.6 Wang pedigree – Y-chromosome haplotypes



List of references

Barbujani G., A. Magagni, E. Minch and L.L. Cavalli Sforza (1997). An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America* **94(9):** 4516-4519.

Bierne, N., S. Launey, Y. Naciri-Graven and F. Bonhomme (1998). Early effect of inbreeding as revealed by microsatellite analyses on *Ostrea edulis* larvae. *Genetics* **148**:1893-1906.

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

Bittles, A. H. (1998). *Empirical estimates of the global prevalence of consanguineous marriage in contemporary societies*. Morrison Institute for Population and Resource Studies, Working paper no. 74. Stanford University

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

Cann, R.L., M. Stoneking and A.C. Wilson (1987). Mitochondrial DNA and human evolution. *Nature* **325**: 31-36.

Casanova M., P. Leroy, C. Boucekkine, J. Weissenbach, C. Bishop, M. Felous *et al.* (1985). A human Y-linked DNA polymorphgism and its potential for estoimating genetic and evolutionary distance. *Science* **230**: 1403-1406.

Cavalli – Sforza L.L. and W.F. Bodmer (1971). *The genetics of human populations.* W.H Freeman Publishers, San Francisco.

Cavalli-Sforza, L.L., P. Menozzi and A. Piazza (1994). *The history and geography of human genes*. Princeton University Press, Princeton.

Chakravarti, A. (1998). Population genetics – making sense out of sequence. *Nature Genetics (supplement)* **21**: 56-60.

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

Chinese Family Planning Commission (1997). *Chinese Family Planning Yearbook 1997.* Family Planning Commission, Beijing.

Chu, J. Y., Huang W., Kuang S.Q. *et al.* (1998). Genetic relationship of populations in China. *Proceedings of the National Academy of Sciences, U.S.A.* **95:** 11763-11768.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution.* **23**: 72-84.

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

Comas, D., F. Calafell, E. Mateu, A. Péréz-Lezuan, E. Bosch, R. Martínez-Arais *et al.* (1998). Trading genes along the Silk Road: mtDNA sequences and the origin of Central Asian populations. *American Journal of Human Genetics* **63:** 1824-1838.

Coulson T.N., J.M. Pemberton, S.D. Albon, M. Beaumont, T.C. Marshall, J. Slate *et al.* (1998). Microsatellites reveal heterosis in red deer. *Proceedings of the Royal Society of London B* **265:** 489-495.

Cooper, G., W. Amos, D. Hoffman, and D.C. Rubenstein (1996). Network analysis of human Y microsatellite haplotypes. *Human Molecular Genetics* **5:** 1759 - 1766.

Dessaint, W. L. (1995-1996). The Lisu, highlanders of the Salaween. *Bulletin of the International Committee on Urgent Anthropological and Ethnological Research* **37-38:** 13-27.

Dikotter, F. (1992). *The discourse of race in modern China.* Stanford University Press, Stanford University.

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

Dorsten, L.E., L. Hotchkiss and T.M. King (1999). The effect of inbreeding on early childhood mortality: Twelve generations of an Amish settlement. *Demography,* **36:** 263-271.

Du, R. and Z.L. Zhao (1981). Percentage and types of consanguineous marriages of different nationalities and regions in China. *National Medical Journal of China* 61, 723-728.

Du, R. and V.F. Yip (1993). *Ethnic Groups In China.* Science Press, Beijing and New York.

Excoffier, L., P.E. Smouse, and J. Quattro (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* **131:** 479-491.

Fairbank, J and L. Reicher (1990). *China: transformation and transition.* Allen & Unwin Australia, Sydney.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **35:** 785-791.

Felsenstein, J. (1989). PHYLIP - phylogeny inference package (version 3.2). *Cladistics* **5:** 164-166.

Gladney, D. C. (1996). *Muslim Chinese: ethnic nationalism in the Peoples Republic.* Harvard University Press, Cambridge, Massachusetts.

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

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

Guo, S. W. and E.A. Thompson (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48:** 361-372.

Hammer, M.F., T. Karafet, A. Rasanayagam, E.T. Wood, T.K. Altheide, T. Jenkins *et al* (1998). Out of Africa and back again: Nested cladistic analysis of human Y-chromosome variation. *Molecular Biology and Evolution* **15(4):** 427-441.

Hartl, D. L. and A.G. Clark (1997). *Principles of population genetics.* Sinauer Associates, Massachusetts.

Hirszfeld, L. and H. Hirszfeld (1918-1919). Essai d'application des méthodes sérologigues au probléme des races. *Anthropologie* **29:** 505-537.

Hopkirk, P. (1980). *Foreign Devils on The Silk Road.* Oxford University Press, Oxford.

Jeffereys A., N. Royle, V. Wilson and Z. Wong (1985). Hypervariable "minisatellite" regions in human DNA. *Nature* **314:** 67-73.

Jobling, M. A. (1995). Fathers and sons: The Y-chromosome and human evolution. *Trends In Genetics* **11:** 449-456.

Jorde, L. B., (1997). Inbreeding in human populations. in *Encylopedia of Human Biology* 2nd edition **5:** 1-13. Academic Press, San Dieago

Karafet, T.M., S.L. Zegura, O. Posukh, L. Osipova, A. Bergen *et al* (1999). Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *American Journal of Human Genetics* **64(3)**: 817-835.

Kayser, M., A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi *et al* (1997). Evaluation of Y - chromosomal STRs: a multicenter study. *International Journal of Legal Medicine* **110:** 125 - 133.

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

Kittles, R., M. Perola, L. Peltonen, A.W. Bergen, R.A. Aragon, M. Virkkunen *et al*, (1998). Dual origins of Finns revealed by Y-chromosome haplotype variation. *American Journal of Human Genetics* **62:** 1171-1179.

de Knijff, P., M. Kayser, A. Caglia, D. Corach, N. Fretwell, C. Gehrig *et al* (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. *International Journal of Legal Medicine* **110:** 134-49.

Leslie, D. D. (1986). *Islam in traditional China: a short history to 1800.* Canberra College of Advanced Education, Canberra.

Lewontin, R.C. (1972). The apportionment of human diversity. *Evolutionary Biology* **6**: 381-398.

Lipman, J.N. (1997). *Familiar strangers: a history of Muslims in northwest China.* Hong Kong University Press, Hong Kong University.

Lucotte, G. and N.Y. Ngo (1985). p49F, a highly polymorphic probe that detects TaqI RFLPs on the human Y chromosome. *Nucleic Acids Research* **13**: 8285.

Malaspina, P., F. Cruciani, B.M. Ciminelli, L. Terrenato, P. Santolamazza, A. Alonso *et al.* (1998). Network analysis of Y-chromosomal types in Europe, Northern Africa and western Asia reveal specific patterns of geographic distribution. *American Journal of Human Genetics* **63:** 847-860.

Markow T., P.W. Hedrick, K. Zuerlein, J. Danilovs, J. Martin, T. Vyvial *et al.* (1993). HLA polymorphism in the Havasupai: evidence for balancing selection. *American Journal of Human Genetics* **53:** 943-952.

Marks, J. (1995). *Human biodiversity: genes, race and history.* Aldine De Gruyter, New York.

Mickalakis, Y. and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* **142**: 1061-1064.

Minch, E., (1997) *MICROSAT*, Stanford University. At: http://human.stanford.edu

Morell, R., Y. Liang, J. Asher, J. Weber, J.T. Hinnant, S. Winata *et al.* (1995). Analysis of short tandem repeat (STR) allele frequency distributions in a Balinese population. *Human Molecular Genetics* **4:** 85-91.

Neel, J.V. and R.K. Ward (1972) Genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by *F*-statistics, including a comparison with the Mkirate and Xavante. *Genetics* **72**: 639-666.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Science USA* **70:** 3321-3323.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia University Press, New York.

Péréz-Lezuan, A., F. Calafel, E. Mateu, D. Comas, R. Ruiz-Pacheco and J. Bertranpetit (1997a). Microsatellite variation and the differentiation of modern humans. *Human Genetics*. **99**:1-7.

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

Péréz-Lezuan, A., F. Calafel, D. Comas, E. Mateau, E. Bosch, R. Martinez-Arais *et al* (1999). Gender–specific migration patterns in central Asian populations revealed by the analysis of Y-chromosome STRs and mtDNA. *American Journal of Human Genetics* **65:** 208-219.

Peterson, A.C., A. Di Rienzo, A. Lehesjoki, A. de la Chapelle, M. Slatkin and N.B. Freimer (1995). The distribution of linkage disequilibrium over anonymous genome regions. *Human Molecular Genetics* **4:** 887-894.

Pritchard, J.K. and N.A. Rosenberg (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**:220-228.

Rahman, Y. A. (1997). *Islam in China.* at: http://www.erols.com/ameen/islchina

Raymond, M. and F. Rousset (1995). An exact test for population differentiation. *Evolution* **49:** 1280 - 1283.

Roewer, L., M. Kayser, K Dieltjes., M. Nagy, E. Bakker, M. Krawczak & P.de Knijff (1996). Analysis of molecular variance (AMOVA) of Y-chromosome specific microsatellites in two closely related human populations. *Human Molecular Genetics* **5:** 1029-1033.

Rousset, F. (1995). GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicalism. *Journal of Heredity* **83:** 239.

Rousset, F. (1996). Equilibrium values of measure of population subdivision for stepwise mutation processes. *Genetics* **142:** 1357 - 1362.

Rousset, F. and M. Raymond (1995). Testing heterozygote excess and deficiency. *Genetics* **140:** 1413 - 1419.

Saitou, N. and. M. Nei (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biological Evolution* **4:** 406-425.

Senior, P.A. and B. Raj (1994). Ethnicity as a variable in epidemiological research. *British Medical Journal* **309:** 327-330.

Shami, S.A., J.C. Grant and A.H. Bittles (1994) Consanguineous marriage within social/occupational class boundaries in Pakistan. *Journal of Biosocial Science* **26:** 91-96.

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

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* **137:** 331-336.

Slatkin, M. (1995). Measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139:** 457-462.

Stoltenberg, C., P. Magnus, A. Skrondal and R. Terje Lie (1999). Consanguinity and recurrence risk of stillbirth and infant death. *American Journal of Public Health* **89:** 517-523.

Takezaki, T. and M. Nei (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144:** 389-399.

Taylor, A.C., W.B. Sherwin and R.K. Wayne (1994). Genetic variability of microsatellite loci in a bottleneck species: the northern hairy-nosed wombat *Lasiorhinus krefftii. Molecular Ecology* **3:** 277-290.

Templeton, A.R. and B. Read (1983). The elimination of inbreeding depression in a captive herd of Speke's gazelle. In: *Genetics and Conservation: A reference for managing wild animal and plant populations.* Addison-Wesley.

Templeton, A.R. and B. Read (1984). Factors eliminating inbreeding depression in a captive herd of Speke's gazelle (*Gazella spekei*). *Zoo Biology* **3:** 177-199.

Templeton, A.R. and B. Read (1998). Elimination of inbreeding depression from a captive population of Speke's gazelle: Validity of the original statistical analysis and confirmation by permutation testing. *Zoo Biology.* **17(2):** 77-94.

Wainscoat, J.S., A.V.S. Hill, A.J. Boyce, J. Flint, J. Hernandez, S.L. Thien *et al.* (1986). Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. *Nature* **319(6053):** 491-493.

Wang, W., S.G. Sullivan, S. Ahmed, D. Chandler, L.A. Zhivotovsky and A.H. Bittles (2000). A genome based study of consanguinity in three co-resident endogamous Pakistan communities. (in press).

Wayland Barber, E. (1999) *The mummies of Ürümchi.* Macmillan, London.

Weber, J.L. and P.E. May (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics.* **44:** 388-396.

Weber, J. L. and C. Wong (1993). Mutation of human short tandem repeats. *Human Molecular Genetics.* **2:** 1123-1128.

Weir, B.S. and C.C. Cockerham (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38(6):** 1358-1370.

Weir, B.S. (1996). *Genetic data analysis II.* Sinauer Associates Inc. Sunderland, Massachusetts.

Weiss, K.M. (1998) Coming to terms with human variation. *Annual Review of Anthropology* **27**: 273-300

Wong, H. M. and A. A. Dajani. (1988). *Islamic frontiers of China.* Scorpion, London.

Workman, P.L., H. Harpending, J.M. Lalouel, C. Lynch, J.D. Niswander and R. Singleton (1973). Population studies on southwestern Indian tribes VI: Papago population structure: a comparison of genetic and migration analyses. In: *Genetic structure of populations.* University of Hawaii Press, Honolulu.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* **15:** 323-354.

Wu, L. (1987). Investigation of consanguineous marriages among 30 Chinese ethnic groups. *Heredity and Disease* **4:** 163-166. [In Chinese].

Yifu, S., ed. (1989). *The Silk Road on land and sea* China Publishing, Beijing.

Zerjal, T., B. Dashnyam, A. Pandya, M. Kayser, L. Roewer, F.R. Santos *et al.* (1997). Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosome analysis. *American Journal of Human Genetics* **60:** 1174-1183.

Zhan, J., W. Qin, Y. Zhou, K. Chen, W. Yan and W. Yu (1992). Effects of consanguineous marriages on hereditary diseases: a study of the Han ethnic group in different geographic districts of Zejiang province. *National Medical Journal of China* **172:** 674-676. [In Chinese].