MICROSATELLITE VARIATION IN HUMAN POPULATIONS OF INDIA

MICROSATELLITE VARIATION
IN HUMAN POPULATIONS
OF THE INDIAN SUBCONTINENT

By

Sujatha Thampi, B.Sc.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (1997)
(Biology)

McMASTER UNIVERSITY
Hamilton, Ontario

TITLE: Microsatellite Variation In Human Populations of the Indian Subcontinent

AUTHOR: Sujatha Thampi, B.Sc. (McMaster University)

SUPERVISOR: Professor R.S. Singh

NUMBER OF PAGES: ix, 184

Abstract

An analysis of microsatellite variation among individuals of Indian origin was performed to compare the genetic diversity between different regions within India that are divided by spoken language and geographical distance. In addition, the relationship between the Indian and other human populations was examined. Six microsatellite loci - CSF1PO, TPOX, TH01, F13A01, FESFPS and vWA were amplified and typed in a total of 149 individuals representing a broad geographical distribution within the Indian subcontinent. Contingency analyses of allele frequency distributions between North and South Indian populations revealed a significant difference at the vWA locus. The lack of significant differences at the other five loci may indicate a considerable amount of gene flow between these two populations or that these polymorphisms existed before the split between Northern and Southern populations. The South Indian population (n = 37) revealed the presence of an allele at the vWA locus that was not detected in either the North Indian (n = 103) or Caucasian (n = 212) populations. The absence of this allele in the North and the presence in the South may indicate a population specific allele and gives support to the hypothesis that India was occupied by an earlier Dravidian population before the more recent arrival of the Aryans who lacked this allele. Genetic distance failed to reveal any significant distances between North and South Indian populations. Phylogenetic analyses, although not significant, indicated that the two populations were not monophyletic. A correlation between genetic and geographic distance between the Northeast, Northcentral, Southeast and Southwest regions of India revealed a correlation coefficient of r = -0.53 which was not significant. The negative correlation is solely due to the fact that the two most geographically closest populations, Salem and Cochin, show the greatest genetic distance

(0.442) between them. This is consistent with the fact that social factors play an important role in the genetic structuring of Indian populations. Genetic distance analyses revealed significant distance values between the Indian, Caucasian and African populations and showed the Indian population to be genetically closer to the Caucasian population. These results are consistent with earlier studies using gene frequency data and mitochondrial DNA. Overall, the results of the genetic distance and phylogenetic analyses showed that six microsatellite loci were able to distinguish between African and non-African populations, but more loci need to be utilized to differentiate between non-African populations.

Acknowledgements

It is not possible to embark on such a project without the support and help of family, friends and teachers. Therefore, it is with the utmost appreciation that I acknowledge all those who made it possible for me to achieve my goal.

I would like to thank my supervisor, Dr. Rama Singh for his support, patience and guidance and I would also like to thank Dr. Richard Morton for his assistance and the time he took to answer all my questions.

I will always be grateful to Fariborz Yazdani for his unselfish willingness to help me out with one problem or another, at any time. I must also thank Robert Dawson for starting me out on the right path. I would also like to thank all past and present members of lab 505 for their support, including Aaron Thomson and in particular Anouk Behara who eased my transition into the lab.

Last but not least, I thank Alberto Civetta and Rob Kulathinal who took the time to listen, advise and who made my time as a graduate student most enjoyable and memorable.

Finally, I would like to thank my family for their continual support, and belief in me.

Table of Contents

Introduction.. 1
The Indian Subcontinent.. 1
Prehistory.. 6
The Indus Valley Civilization.................................... 7
Downfall of the Indus Valley Civilization....................... 8
The Aryan Arrival.. 8
The Dravidian Peoples... 10
Later Invasions.. 11
Genetic Surveys of the Indian Population....................... 13
Microsatellite Loci.. 17
Previous Human Population Studies based on Microsatellite Loci........ 21
Objectives of the Present Study................................. 23

Materials and Methods... 25
Sampling Procedure.. 25
DNA Extraction.. 25
PCR Amplification of Six Microsatellite Loci................... 29
Polyacrylamide Gel Electrophoresis.............................. 31
Silver Staining Polyacrylamide Gels............................. 35
Heterozygosity... 38
Hardy-Weinberg Equilibrium...................................... 38
Contingency Analysis.. 41
Descriptive Statistics.. 46
Genetic Distance.. 46
Phylogenetic Analysis... 47
Correlation.. 47
Fst.. 48

Results.. 49
Heterozygosity... 49
Hardy-Weinberg Equilibrium...................................... 49
Contingency Analysis and Descriptive Statistics................ 52
Genetic Distance.. 108
Phylogenetic Analysis... 117
Correlation between Genetic and Geographic Distance of the Four Regions...... 127
Fst.. 127

Discussion... 134
Hardy-Weinberg Equilibrium...................................... 134
Relationship between North and South India and Other World Populations........ 135
Relationship of the Indian Population to Other World Populations............. 140
Relationship of Populations Within India........................ 142
The Use of Microsatellites in Population Genetic Studies....... 144
Future Work... 149

Conclusion.. 151

Appendices... 153
Appendix A - Protocols.. 153
Appendix B - Solutions.. 164
Appendix C - Data... 173

References.. 179

List of Tables

Table 1: Geographic distribution of samples.. 28
Table 2: Locus-specific information.. 30
Table 3: PCR product sizes and positive control genotypes.................................... 32
Table 4: Source and sample sizes of populations.. 39
Table 5: Observed and expected heterozygosities.. 50
Table 6: Test of Hardy-Weinberg Equilibrium for the Indian population.................... 51
Table 7: Contingency analysis for North and South India... 53
Table 8: Descriptive statistics for North and South India.. 66
Table 9: Contingency analysis for India vs. other human populations at the CSF1PO locus.. 70
Table 10: Contingency analysis for India vs. other human populations at the TPOX locus.. 73
Table 11: Contingency analysis for India vs. other human populations at the TH01 locus.. 76
Table 12: Contingency analysis for India vs. other human populations at the F13A01 locus.. 79
Table 13: Contingency analysis for India vs. other human populations at the FESFPS locus.. 82
Table 14: Contingency analysis for India vs. other human populations at the vWA locus.. 85
Table 15: Summary of contingency analysis between India and other populations.....
Table 16: Descriptive statistics for India and other populations.................................. 88
Table 17: Contingency analysis for populations within India at the CSF1PO locus..... 92
Table 18: Contingency analysis for populations within India at the TPOX locus.......... 95
Table 19: Contingency analysis for populations within India at the TH01 locus........... 98
Table 20: Contingency analysis for populations within India at the F13A01 locus....... 101
Table 21: Contingency analysis for populations within India at the FESFPS locus...... 104
Table 22: Contingency analysis for populations within India at the vWA locus........... 107
Table 23: Descriptive statistics for populations within India...................................... 109
Table 24: Genetic distances between four populations using six loci......................... 110
Table 25: Genetic distances between five populations using six loci......................... 111
Table 26: Genetic distances between six populations using four loci......................... 112
Table 27: Genetic distances between seven populations using four loci.................... 113
Table 28: Trend between genetic distance and % of loci that have significant differences in allele frequency distributions from the Indian population at six loci.. 115
Table 29: Trend between genetic distance and % of loci that have significant differences in allele frequency distributions from the Indian population at four loci.. 116
Table 30: Genetic distances between populations within India.................................. 118
Table 31: Correlation between geographic and genetic distance within India............. 128
Table 32: Fst values for six loci when four populations are pooled............................ 129
Table 33: Fst values for four loci when six populations are pooled............................ 131
Table 34: Trend between Fst and % of populations with significant differences in allele frequency distributions from the Indian population at six loci............ 132
Table 35: Trend between Fst and % of populations with significant differences in allele frequency distributions from the Indian population at four loci........... 133

Introduction

The Indian Subcontinent

The Indian civilization is one of the oldest, having a history stretching back more than 4000 years and is considered to be one of the cradles of civilization alongside Egypt and Mesopotamia (Tammita-Delgoda, 1995). Bounded to the north by the Himalayan mountain chain, and bordered on the west and east coast by the Arabian Sea and the Bay of Bengal, India is virtually isolated from the rest of Asia. As a result, the Indian land mass is often referred to as a subcontinent. From north to south, the peninsula narrows to a tip at Cape Comorin and measures almost 3000 kilometers. It is about the same distance from east to west, and has a shore line of about 7000 kilometers making it the seventh largest country in the world (Deshpande, 1994; Goren, 1993). See **figure 1** for a map of India.

The Himalayan mountain region to the north stretches west and east forming a range of hills and mountains which in the past, has acted as a barrier to separate India from the rest of Asia. The hills of the Northeast are steep and densely forested, forming the Arakan range. To the Northwest, lies the Karakoram range and the Hindu Kush mountains, where several passes have served as passageways for generations of foreign invaders and migrants (Tammita-Delgoda, 1995). The mountainous northern regions, and the oceans to the east and west, have isolated the people and cultures of India, but has not prevented the movement of people into the subcontinent.

India contains 1/7th of the worlds population, and has one of the most diverse group of people. In 1991, the population was 844 million, making it the worlds second most populated country next to China (Goren, 1993; Cavalli-Sforza *et al.*, 1994). Over thousands of years, several groups of people have migrated to the subcontinent, and the intermixture of these people has produced a degree of uniformity. **Figure 2** shows a timeline of the important historical events that

Figure 1: Map of India and surrounding countries.

India and Surrounding Countries

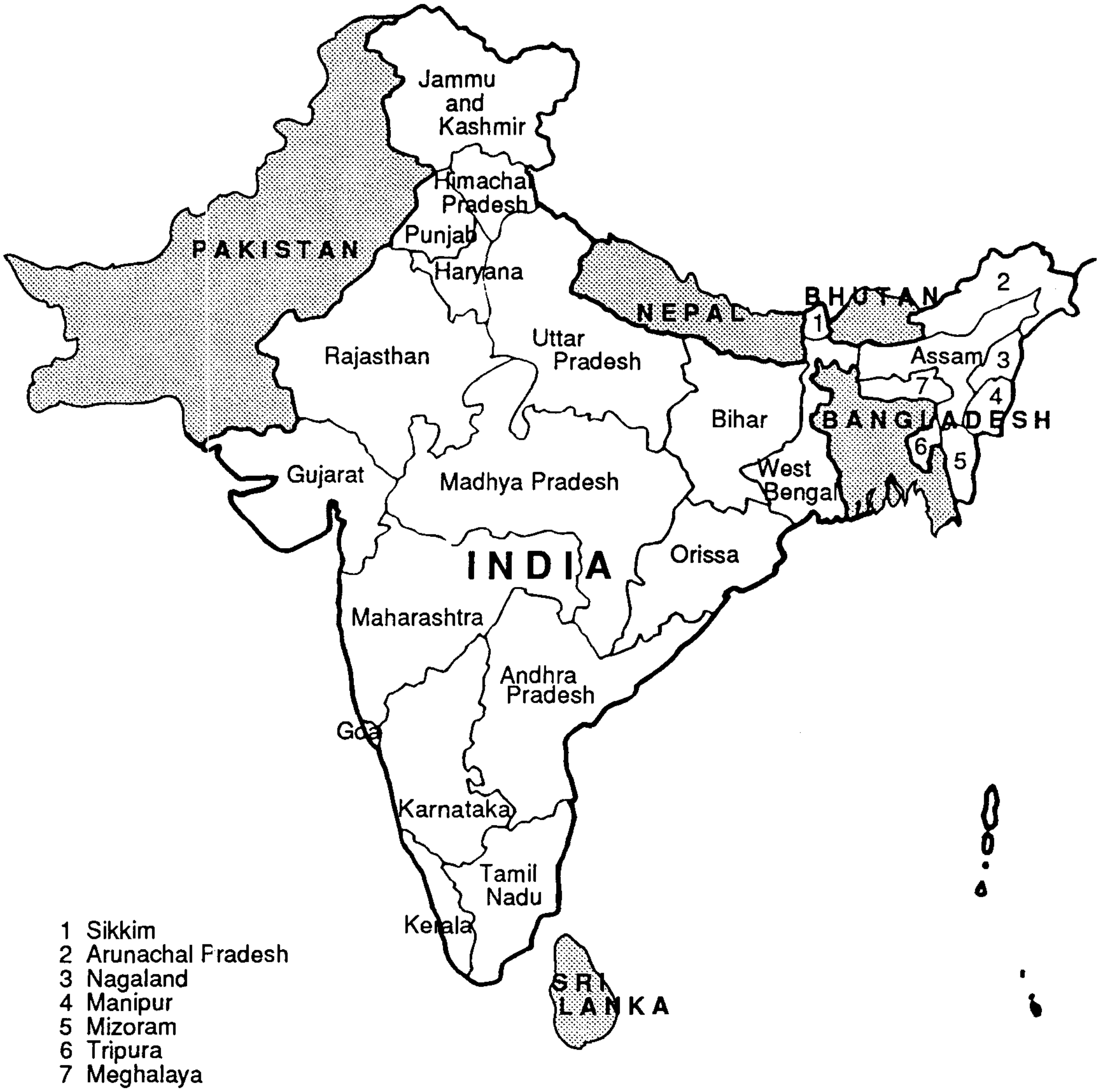


Figure 2: Timeline of significant historical events that impacted on the human diversity of India.

	Dravidians in India
3000 B.C.	Rise of the Indus Valley civilization
2000 B.C.	
1750 B.C.	Decline of the Indus Valley civilization
1500 B.C.	Aryan tribes enter India
1000 B.C.	
520 B.C.	Achaeminid empire
185 B.C.	Scythians - Central Asian tribes enter India
0	
320 A.D.	Gupta Dynasty
540 A.D.	White Huns - Mongoloid tribes enter India
1000 A.D.	Turks, Afghans, Persians enter India
1192 A.D.	Delhi Sultanate - Turkish rule
1498 A.D.	European contact begins
1576 A.D.	Mughal Empire
1700 A.D.	Colonialism attempts begin

impacted on the human diversity of India. However, despite this uniformity, many differences originating in the past, such as physical, cultural, religious and linguistic features, brought in by the different migrants still exist today and may contribute to the vast genetic diversity within India.

Prehistory

According to archaeological evidence, the earliest known humans first migrated to the Indian subcontinent following the second Ice Age, between 400 000 and 200 000 years ago. It is thought that some of these primitive peoples crossed the Hindu Kush mountains into the area that is now Pakistan, while others may have sailed to Southern India from regions of Eastern Africa (Franda and Franda, 1994; Tammita-Delgoda, 1995).

Although no human remains exist, crude stone tools have been discovered that date back to the Old Stone Age(Paleolithic). Two distinct tool industries have been recognized, based on different styles of tool making. These two cultures occupied different regions of India, and are thought to have arisen at the same time. The Sohan industry was discovered throughout North-western India, especially in the Punjab region around the Soan River Valley. The Madras industry was discovered to be widely distributed throughout Southern and Central India (Begley, 1994; Tammita-Delgoda, 1995).

Following the Paleolithic era, the Middle Stone Age(Mesolithic) saw the development of crude pottery and implements made from bone and flint. Evidence of Mesolithic cultures are found from 10 000 years ago (Cavalli-Sforza *et al.*, 1994; Tammita-Delgoda, 1995).

The New Stone Age(Neolithic) culture seemed to flourish in Mesopotamia, Egypt, and Persia between 9000 and 5000 B.C.. However, the Neolithic culture only arrived later in India around 4000 B.C.. The former nomadic ways of hunting and gathering were abandoned in favor of agricultural settlements and the domestication of animals. The first series of small agricultural settlements were discovered in the Northwest hills of Baluchistan, dating from 3712 and 3688 B.C.. From about 3000 B.C., similar agricultural settlements appeared throughout India, concentrated mostly around the Indus River Valley (Begley, 1994; Tammita-Delgoda, 1995).

The Indus Valley Civilization

The origins of the Indus Valley or Harappan civilization dates from between 3000 and 2500 B.C., and is thought to have lasted almost 1000 years. This civilization is thought to have developed from a pre-Harappan culture. As discussed above, several small agricultural settlements were discovered around the Indus River Valley. Among one of the oldest sites is Amri, dating between 3320 and 3600 B.C. (Tammita-Delgoda, 1995; Cavalli-Sforza *et al.*, 1994).

The Indus Valley Civilization was based on the two largest cities of Harappa and Mohenjo-Daro, which were located 400 miles apart and at their peak, is thought to have been inhabited by 35 000 people. These two cities were located at either end of an area that was occupied by the Indus culture. The size of the Indus River Valley civilization was twice the size of the Old kingdom of Egypt, and the Indus culture is thought to have occupied an area just under 1.5 million square miles. Remains of more than 100 cities, towns and villages have been located (Tammita-Delgoda, 1995; Begley, 1994).

The Harappan civilization was one of the most advanced of ancient times. Excavated cities revealed brick houses that were two or three levels high, and well laid out roads and drainage systems. Material remains of the civilization that have been unearthed, include tools made of copper, bronze and stone, gold and semi-precious jewelry, cotton clothing, standardized weights and sculptures. Furthermore, small seals with Harappa script, that have yet to be deciphered has also been discovered in Mesopotamia, indicating that a trading system probably existed between these two civilizations (Franda and Franda, 1994). Biometric measurements and comparisons of cranial series of Harappans to series from sites dated to about the same time reveal similarities between the Harappans and Tepe Hissar and Sakkara civilizations which were based in Iran and Egypt respectively. This suggests a possible relation to Mesopotamian civilizations (Dutta, 1984).

Downfall of the Indus Valley Civilization

No conclusive evidence is available to explain the demise of the Indus Valley Civilization, which is thought to have occurred between 1750 and 1500 B.C., although several theories exist. Archaeologists believe that the decline of the Harappan civilization was the result of a series of natural disasters which may have eventually led to social and economic disturbances (Tammita-Delgoda, 1995; Cavalli-Sforza *et al.*, 1994; Begley, 1994).

Around 1700 B.C., the Indus River changed course which resulted in a series of disastrous floods. These long lasting and widespread floods most probably ended the agricultural system upon which the Harappan civilization relied on for sustenance. By 1650 B.C., signs of chaos were evident in the construction of houses that were cramped and disorganized, very unlike the well planned houses and cities of the past. It is believed that by 1500 B.C., the Indus Valley Civilization had completely disintegrated into fragmented communities, sometimes referred to as post-Harappan (Tammita-Delgoda, 1995; Cavalli-Sforza *et al.*, 1994).

Another theory, and perhaps the most prevalent to explain the disappearance of the Indus Valley Civilization, has often been attributed to the Aryan invasions from the Northwest, but, no evidence for a violent take-over has been found. However, the eventual demise of the Harappan civilization does seem to coincide with the arrival of the Aryan pastoral nomads, who came from the steppes of Central Asia (Tammita-Delgoda, 1995; Cavalli-Sforza *et al.*, 1994; Begley, 1994; Franda and Franda, 1994).

The Aryan Arrival

The Aryan people were a group of semi-nomadic tribes who originally inhabited the steppes of Central Asia, in particular the region between the Caspian Sea and the Black Sea. Around 2000 B.C., these tribes left their homeland in Southern Russia for reasons that were not known. However, it has been speculated that plague, famine or natural disasters may have been causes for their emigration (Tammita-Delgoda, 1995). From the steppes of Central Asia, the Aryans moved in every direction. Some tribes moved into Western Asia and onto the Iranian plateau, before moving east again, towards the Indian subcontinent around 1500 B.C.. It is thought that they advanced through the passes found in the Northwest Hindu Kush mountains and into Punjab. Although no archaeological evidence has yet been identified with absolute certainty, that would mark the arrival of the Aryans, they may be associated with the appearance of Painted Gray Ware dated between 1000 and 500 B.C. (Embree, 1988; Cavalli-Sforza *et al.*, 1994; Begley, 1994).

Two important aspects of Indian culture are associated with the migratory Aryan people. The first, is the growth and spread of the Indo-European language family in India, out of which developed the Indo-Aryan group of languages. These include classical Sanskrit, as well as many of the modern languages predominantly spoken in Northern India such as Hindi, Bengali and Marathi. The other aspect of Indian culture attributable to the Aryan people is the vast body of Vedic literature, which are the oldest texts written in Sanskrit. Originally transmitted orally, the Vedas are a collection of hymns, poems, prayers, philosophic ideas and instructions that have been preserved almost unchanged, for 3000 years (Embree, 1988).

Knowledge of the Aryan people comes not from archaeology, but from the Vedas. Despite being semi-nomadic in nature, in certain respects, they were more advanced than many of the more sophisticated cultures of the time. They used iron, and are believed to have been the first to tame and harness the horse to a chariot, a revolutionary new development (Tammita-Delgoda, 1995). Also mentioned in the Vedas was a system of *varnas,* or classes which described a strictly stratified Aryan society, very unlike the Harappan culture. From this system of social structuring, evolved the caste system. The four main varnas in the caste system included Brahmins(priests), Kshatriyas(political rulers or warriors), Vaishyas(traders and cultivators) and Shudras(artisans). Even today, the Vedas and the caste system remain central to the Indian socioreligious system of Hinduism (Embree, 1988; Franda and Franda, 1994).

Over the next few centuries, the post Harappan and Aryan cultures are believed to have gradually fused in Northern India as the Aryan people gradually expanded eastward onto the Gangetic plain. Between 1000 and 500 B.C., growth of new towns, settled agriculture, trade and various changes in the status of tribal rulers took place. These developments mainly took place in the Gangetic heartland. Elsewhere at this time, different social, cultural and linguistic communities formed in Eastern and Southern India. This was particularly true in South India, whose languages were not Indo-Aryan, but belonged to the Dravidian linguistic family. Despite these differences, there was probably some interaction between these diverse ethnic and linguistic groups throughout the centuries as some customs and traditions were shared (Embree, 1988, Tammita-Delgoda, 1995).

The Dravidian Peoples

The word "Dravidian" is sometimes applied as a racial category to describe the darker-skinned inhabitants of South India and Sri Lanka. This word is also used to refer to the family of languages spoken in Southern India and it is also used as a cultural concept to describe the particular forms of Hindu art, architecture and literature that developed in this region (Radhakrishnan, 1994). Furthermore, this term is also used to describe cultural beliefs and practices, including a distinctive kinship system (Cutler, 1988).

The origins of the Dravidian people are uncertain, however, the widely accepted theory is that they probably arrived from the region around Western Iran sometime before the Aryans reached India (Cavalli-Sforza *et al.*, 1994). Linguistic evidence suggests that the Dravidian culture was previously spread over much of Northern India, since today, there are more than 2 million speakers of Dravidian languages in Pakistan, Afghanistan and the border areas of India (Radhakrishnan, 1994). It has also been speculated that the Dravidian civilization may have arisen out of the Harappan culture, however, linguistic evidence does not seem to support this theory since Harappa script still remains undeciphered (Cutler, 1988). To explain the origins of the division between the Indo-Aryan linguistic groups in the North and the Dravidian languages of the South, it has been proposed that the invading Aryans who spoke Indo-European languages, on arrival pushed south the darker-skinned inhabitants of the north, whom they refer to as *dasas* in the Vedas (Tammita-Delgoda, 1995; Franda and Franda, 1994).

The Dravidian family of languages is predominantly spoken in South India. However, some inhabitants of Sri Lanka, Baluchistan and isolated regions of Central India speak Dravidian languages. The four main branches of the family are Tamil, Telugu, Kannada and Malayalam. Tamil is the oldest, and most highly developed. It is the first language to develop a literature of its own, and in the spoken form, is thought to be older than Sanskrit. Connections linking the Dravidian family to other linguistic families have been sought, but none have been established (Emeneau, 1994; Cutler, 1988).

The persistence of a very large number of speakers of Dravidian languages in Central and South India, despite the domination of the Indo-Aryan languages of the north, may be an indirect indication that their genetic identity was not seriously altered by the arrival of the Aryans. Certainly, there are still many social and cultural distinctions between North and South India that still remain today (Cavalli-Sforza *et al.*, 1994).

Later Invasions

The period between 600 B.C. and 300 A.D. was a time of intense cultural and social change in India, marked by a number of invasions and the rise and fall of various empires. Throughout this period, the Northwest regions of India had repeated contacts with the empires of the Iranian plateau and by 520 B.C., the area west of the Indus River became part of the Achaeminid empire. In 326 B.C., Alexander the Great conquered the Achaeminid empire and advanced as far east as Islamabad, however, he departed India shortly after his entry, and his death in 323 B.C. prevented further Macedonian invasions (Majumdar, 1994).

Around 324 B.C., North India came under the control of Chandragupta Maurya, and later, under his grandson Ashoka, the Maurya dynasty was established. Maurya power extended over much of India making it the country's first large empire. At about this time, kingdoms were established in South India, although these Dravidian kingdoms were known to have existed from at least the 1st century B.C..

After the decline of the Maurya dynasty around 185 B.C., North India came under the rule of local dynasties as well as various intruders from Northwest Asia, including the Scythians. Known as Sakas in Indian history, these nomadic tribes from Central Asia established the Kushan dynasty lasting from 78 to 200 A.D. and were best known for spreading Buddhism throughout Northern India.

The Gupta dynasty was established in North India around 320 A.D. and lasted till 540 A.D.. This time is generally considered to be ancient India's classic period when Indian architecture, sculpture, painting, dance and music flourished. However, the Gupta dominance was weakened, both by internal struggles, and by invasions from the Northwest, by an East Asian people known as the White Huns. The following centuries saw kingdoms rise and fall in both North and South India.

Islam first entered India in 711 A.D., when an Arab general from the Eastern provinces of the Umayyad empire conquered Sind, in what is now Pakistan. However, the Arab conquest in Sind had little influence on the rest of India. The chief Muslim conquests of India came from Central Asian converts to Islam. In particular, the Turks, Afghans, Persians and Mongols began to enter the subcontinent around 1000 A.D..

The period of Turkish rule that lasted from 1192 to 1526 A.D. is referred to as the Delhi Sultanate. Following this reign, the Mughal empire founded by Babur in 1526 began control, lasting till 1707. The Mughal empire is associated with the most brilliant cultural achievements in India. The reasons for the decline of the Mughal empire are not known, however, weakness of the later rulers, religious tensions and the lack of control over their vast territories may have caused the downfall.

Extensive European contact with India began in 1498 when Vasco de Gamma, a Portuguese navigator landed on the Southwest coast. Both the Portuguese and the Dutch attempted to colonize India during the 16th century, but neither proved strong enough to rival the naval power of the French and British. Thus, it would be the British that would finally establish colonial rule in the late 18th century, lasting until 1947, when India would gain its independence (Majumdar, 1994; Franda and Franda, 1994; Embree, 1988; Tammita-Delgoda, 1995; Cavalli-Sforza *et al.* , 1994).

Genetic Surveys of the Indian Population

The people of India exhibit a great deal of genetic diversity, much of which has been attributed to the large scale admixture of people from West Asia, East and Central Asia, and to a lesser extent, people from Africa. Furthermore, the unique social customs of India has resulted in a large number of endogamous populations (Majumder and Mukherjee, 1993). Previous studies addressing the genetic variation of the subcontinent were based mainly on gene frequency data. The classic genetic markers that have been used include, immunoglobins, MHC types, blood groups, allozymes, and more recently, mitochondrial DNA and minisatellites.

In particular, two main types of studies have been conducted that address the genetic variation of the Indian population. The first type, investigates patterns of genetic similarity between the Indian population and neighboring countries to account for the historical migrations and subsequent admixture of people from West, Central and East Asia. The second type of study looks at the genetic similarity of various populations within India by comparing geographic and/or ethnosocial(caste, religion, tribe) background.

Relationship of Indians to Other World Populations

Using gene frequency data from 10 polymorphic protein loci, it was shown that South Indian castes were found to be very close to the Sinhalese of Sri Lanka. This similarity has been ascribed to past large scale migrations of people from Tamil Nadu to Sri Lanka. South Indian tribal groups were found to be genetically close to tribal Veddahs of Sri Lanka and genetically unrelated to the tribal groups of Malaya and the aboriginal groups of New Guinea and Australia (Roychoudhury, 1983; Walter, 1986). Furthermore, non-tribal Dravidian people of South India were shown to be genetically unrelated to Australian aborigines (Cavalli-Sforza *et al.*, 1994; Sanghvi, 1976).

Using data from 4 polymorphic loci(Tf, Gc, Gm and Km), East Indian populations have been shown to be more similar to the East Asian populations of Nepal, Japan and China (Walter, 1986). Roychoudhury (1977) using gene frequency data from 10 polymorphic loci, demonstrated that East Asians in general, show a closer affinity to North Indians rather than South Indians.

Protein and allele frequency data from 10 polymorphic protein loci revealed that the Indian population, represented by North Indians, Bengalis from West India and a South Indian tribe was genetically closest to East Asians and farthest away from Africans with the Caucasians being at the intermediate level (Roychoudhury, 1977). A similar conclusion was reached earlier using serum protein polymorphism's (Walter, 1971). However, later studies showed that caste groups in India were found to be genetically more closer to Iranians and Afghans than to the East Asian populations of Malaya and China (Roychoudhury, 1983). Using 18 protein and blood group loci, it was demonstrated that Indian populations were genetically closer to West Asian populations than to neighboring East Asian populations (Roychoudhury and Nei, 1985).

Nei and Roychoudhury (1993) used gene frequency data of 29 polymorphic loci, to examine the relationships between 24 representative populations from around the world. The Indian population was represented by a group from Punjab. Phylogenetic analyses using the Neighbor-Joining method of construction revealed that the first split occurred between African and non-African populations which was supported by 100% of the bootstraps. The second major split was between Caucasian populations and other non-African populations and the third major split occurred between Native American and Greater Asian populations which included East Asians. The Indian population clustered with the Caucasian populations. However, if a 95% bootstrap value is used as the statistically significant level, the Caucasian cluster which included European Caucasians, Indians and Iranians was not significant. This was thought to reflect the probable admixture of the Indian and Iranian populations with East Asian populations in the past. When these two groups were removed from the phylogenetic analyses, the clustering of the European Caucasians had a bootstrap value of 100%. Previous studies have demonstrated that Indian populations cluster most closely with West Asian populations including Turks, Iranians, and Lebanese and the Indians and West Asians form a Caucasian cluster when compared with Asians from further north and east (Cavalli-Sforza *et al.*, 1994).

Mountain *et al.* (1995) using mitochondrial DNA sequences showed that the Indian population clusters with the European and Chinese, connected with relatively short branches thus indicating a recent common ancestor for these groups.

Recently, populations from Assam were included in studies using minisatellites to asses genetic variation (Deka *et al.*, 1991; Balazs *et al.*, 1992). These studies reveal that groups of alleles at one or more loci have frequencies that are different among some of the populations, and therefore could be used to differentiate one group from another. Thus, as more data from different populations are gathered, the high heterozygosity values at minisatellite loci, as compared to protein and blood group loci may make it possible to study the origins and/or migration patterns of human populations based on frequency distributions at different loci.

In general, the patterns of genetic similarity that are revealed between the Indian populations and neighboring countries, seem to correlate well with the historical accounts of migration, and subsequent integration of West and East Asian populations. Eastern Indian populations tend to genetically closer to East Asian populations as a result of admixture with Eastern populations. Generally, many caste and tribal populations of Northern, Western and Central India show closest affinities with the people of West Asia, while South Indian populations appear to stand apart from the other regions. This may be due to the restriction of the incoming migrants to the northern half of India. Finally, these studies reveal that the Indian population as a whole seems to be genetically closer to West Asian populations than to East Asian, and farthest from African populations (Majumder and Mukherjee, 1993).

Genetic Relationships Within India

Several genetic surveys of India have revealed differences between Northern and Southern populations. Gene frequency data from 10 polymorphic loci show that South Indian populations stand apart from the inhabitants of the rest of India, while populations from the other regions of India are genetically closer to each other (Roychoudhury, 1977). This is in agreement with the sharp linguistic separation of South India with the rest of India and may support the theory of Indo-European speaking migrants, pushing south the Dravidian speaking populations that may have originally inhabited Northern India.

Previous gene frequency studies have also indicated that there is some genetic differentiation between Dravidian and Indo-European language speakers. However, although Dravidian speakers cluster together in gene frequency studies, their branches are not that far removed from Indo-European speakers indicating that there is at least some degree of admixture (Cavalli-Sforza *et al.*, 1994).

Analysis of mitochondrial DNA sequence variation among individuals of Indian origin revealed that there was a significant difference between Northern and Southern populations at 4 out of the 54 variable sites that were detected (Behara, 1995 unpublished data).

Numerous studies have been conducted to examine the genetic relationships between ethnosocial categories with respect to historical events and/or geographical distance. These studies reveal that in certain geographic regions of India, ethnohistory seems to be an important determinant of genetic affinity (Mukherjee, *et al.*, 1979; Kamboh, 1984; Das *et al.*, 1986). However, in other regions, no clear pattern is observed (Saha *et al.*, 1992; Chakraborty, 1986). Similarly, geographic proximity seems to be an important determinant of genetic affinity within some ethnosocial categories (Mukherjee *et al.*, 1979), while in some ethnosocial categories, no geographical effect is discernible (Cavalli-Sforza *et al.*, 1994; Papiha *et al.*, 1982). Thus, it appears that it is not always possible to identify either geographical distance or past ethnic or social connections as the most important determinant of genetic similarity or differentiation within India. The genetic markers that have been used thus far, have yet to reveal any clear patterns of genetic variation and affinities within the Indian subcontinent (Majumder and Mukherjee, 1993).

Microsatellite Loci

Microsatellite loci consist of tandemly repeated DNA segments that are between 2 and 5 nucleotides in length. They are sometimes referred to as short tandem repeats (STR), or simple sequence repeats (SSR) (Edwards *et al.*, 1991; Tautz and Renz, 1984). They are highly polymorphic and are found in large numbers, relatively evenly spaced throughout the genome (Edwards *et al.*, 1991). It has been estimated that there are approximately 500 000 microsatellite loci in the entire human genome (Ashley and Dow, 1994).

The high rate of molecular evolutionary change makes microsatellite loci particularly useful for studying genetic differentiation and phylogenetic relationships among closely related taxa, such as human populations, although they are less useful for interspecific reconstruction (Bowcock *et al.*, 1994). By studying human pedigrees, microsatellites have been estimated to mutate at a rate between 10^{-5} and 10^{-3} mutations per gamete per generation (Edwards *et al.*, 1992; Bowcock *et al.*, 1994). The tree produced for humans using microsatellite data shows strong geographic clustering (Bowcock *et al.*, 1994), as opposed to the geographically scrambled mitochondrial haplotype trees

(Cann *et al.*, 1987). This is due to the high rate of mutation at microsatellite loci, thus providing a more recent view of evolutionary events compared with mitochondrial DNA. Thus, the use of microsatellite data in the study of human populations is thought to be complementary to mitochondrial studies (Ruvolo, 1996).

Advantages of Microsatellites over Minisatellites

Minisatellite loci consist of tandemly repeated DNA segments that are between 10 and 15 nucleotides in length. The advantages of using microsatellites over minisatellites in population genetic studies are the relative ease of scoring gels and the technique.

Microsatellite and minisatellite alleles are inherited in a Mendelian fashion and are co-dominant. However, standard multilocus DNA fingerprinting examines variability at many minisatellite loci simultaneously, therefore, allelic relationships among bands are generally unknown, and genotypes at specific loci cannot be determined. The presence of many bands of unknown loci specificity, makes it difficult to evaluate relationships between samples that are run on different gels (Bruford and Wayne, 1993; Ashley and Dow, 1994). Because of the large number of bands that cannot be identified individually, binning protocols are used with minisatellite loci, which reduces the statistical power of analysis (Budowle *et al.*, 1994).

The polymerase chain reaction of DNA (PCR; Saiki *et al.*, 1983) is utilized to amplify microsatellite loci, since the PCR products are small, ranging from 100 to 500 base pairs. Minisatellite loci are larger, thus the technique of Southern blotting and hybridization with labeled probes is used. PCR provides many advantages over Southern blotting, in that, the technique is faster, and smaller amounts and a lower quality of DNA may be used (Ashley and Dow, 1994).

The Use of Microsatellite Data in Population Genetic Analysis

The genetic distance between two populations gives a relative estimate of time that has passed since the populations were together. Recently, several different estimates of genetic distance measures based on microsatellite data have been developed. These distance measures incorporate different models of evolutionary change for microsatellite variation (Bowcock *et al.*, 1994; Goldstein *et al.*, 1995; Slatkin, 1995; Shriver *et al.*, 1995). The evolutionary process of microsatellite variation is still being investigated, along with estimates for their basic parameters, thus, the understanding of the mutational process is essential before observed variation and genetic distance or population substructure can be inferred.

An empirical study that examined human pedigrees by Weber and Wong (1993) demonstrated that most mutations at microsatellite loci involve the gain or loss of a single repeat unit. The mechanism by which microsatellites mutate are still uncertain. However, two different types of mechanisms have been proposed. The first mechanism involves unequal sister chromatid exchange during meiosis (Weber and Wong, 1993). This type of mechanism produces larger changes in allele size and is used to explain the mutational patterns observed at minisatellite loci or at dinucleotide loci.

The second type of mechanism and the most predominant at microsatellite loci is strand-slippage replication (Levinson and Gutman, 1987). *In vitro* experiments of synthesis of simple sequences that proceed via slippage reactions have led Schlotterer and Tautz (1992) to postulate that stand slippage occurs primarily during lagging strand synthesis during DNA replication. *In vitro* observations suggest that free DNA ends may be necessary for slippage to occur. This mechanism may involve the slippage of the newly synthesized DNA strand upon dissociation of a polymerase complex, producing a bulge. If such a free end would lie within a simple sequence region, it would be possible that slippage can occur before the polymerase complex would be reestablished during lagging strand synthesis. The newly synthesized strand would then contain a bulge which would need to be repaired. DNA repair mechanisms would either remove, or lead to the elongation of a repeat (Schlotterer and Tautz, 1992). This type of mechanism causes small shifts in allele size which can be forward and backward in nature.

The Infinite Allele Model (IAM) of mutation and the Stepwise Mutation Model (SMM) have been proposed to model the above two mechanisms of mutation.

In allozymes, most mutations give rise to a new distinguishable allele in the population. These types of mutations are accounted for by the IAM of mutation. Unequal sister chromatid exchange during meiosis can be modeled by the IAM since large changes in allele size occur and reverting to the original allele would be very difficult and would depend on allele size.

Under the SMM, alleles can only mutate by the gain or loss of one repeat unit. Stand-slippage replication which has been proposed as the predominant mechanism that causes mutations at microsatellite loci that seem to involve the gain or loss of a single repeat unit can be reasonably approximated by the SMM.

However, empirical data suggest that the mutational process may not be exclusively "one step" as modeled by the SMM, therefore, suggesting that sometimes more than one repeat can be gained or lost (DiRienzo *et al.*, 1994; Garza *et al.*, 1995). Thus, the Two Phase Model (TPM) of mutation was developed, incorporating the mutational process of the SMM, but allowing for mutations of larger magnitude to occur (DiRienzo *et al.*, 1994).

In addition to the uncertainty of the mechanisms and models of mutation at microsatellite loci, there appears to be a bias in the mutation rate. Empirical data on interspecific variation support the existence of size constraints on the alleles at microsatellite loci (Bowcock *et al.*, 1994; Garza *et al.*, 1995). This may cause some concern when estimating the time since separation between two populations. When alleles in two populations reach their maximal and minimal values, the allele sizes may start to decrease and increase respectively, thus increasing the number of homoplasies (Murray, 1996). Under the IAM of mutation, no homoplasies exist, however, under the SMM, homoplasies are present. Therefore, under the SMM, the average size difference of alleles between two populations may not necessarily reflect the time of divergence between them and may lead to the overestimation of similarities among populations (Ruvolo, 1996). Template stability may also affect the mutation rate at microsatellite loci, as was demonstrated in an *in vitro* study by Schlotterer and Tautz (1992). This study suggests that repeat length and base composition affect the mutation rate of microsatellite loci. Furthermore, empirical evidence from a comparison of human and primate microsatellite repeats showed than humans have longer loci than those found in other primates, thus, Rubinsztein *et al.*, (1995) speculate that there may be a mutational bias in humans towards an increase in length.

Thus far, simulation and observational studies have been conducted in order to better understand the mutational process taking place at microsatellite loci. These studies indicate that microsatellite loci between 3 and 5 base pairs in length are consistent with the SMM (Shriver *et al.*, 1993), while the evolution of microsatellites that are 1 to 2 base pairs in length can be best explained by the TPM (DiRienzo *et al.*, 1994). It has been recommended that different size classes of microsatellites should not be treated as a single homogeneous data set, as they have different models of evolutionary change. Furthermore, using estimates of genetic distance and population substructure that assume one particular model of the evolutionary process on data that incorporate another model of mutation may yield confusing results (Murray, 1996).

Previous Human Population Studies Based on Microsatellites

Microsatellites are used in a wide range of applications in genetics. They were originally used for genetic mapping and have also been used for linkage analysis in association with disease susceptibility genes (Weissenbach *et al.*, 1992). They have also been utilized to address questions regarding paternity, kinship and individual identification (Queller *et al.*, 1993; Trabetti *et al.*, 1993; Hammond *et al.*, 1994).

More recently, microsatellites are being used to study the genetic relationships between and within human populations.

African and Non-African Split

Many phylogenetic analyses of human populations have suggested that Africans are the most diverged from any other human populations (Nei and Roychoudhury, 1993; Cann *et al.*, 1987). Most of these studies have used blood group or protein loci and mitochondrial DNA as markers.

Using two sets of microsatellite loci, each containing 25 and 8 loci respectively, Nei and Takezaki (1996) were able to demonstrate that the Africans were the first group of people to split from the rest of the human populations. In addition, the tree was rooted using chimpanzee as the outgroup. The root was located between the African and non-African populations. The data set containing 25 microsatellite loci gave statistically significant bootstrap values of 97% and 99% for both the interior branches connecting the chimpanzee to African and non-African populations respectively. The second data set containing 8 microsatellite loci did not reveal significant bootstrap values. This is thought to be due to the fewer number of loci used (Deka *et al.*, 1995; Nei and Takezaki, 1996).

Phylogenies based on mitochondrial DNA display deep African branches indicating a greater diversity for African populations. This points to evidence for an African origin of human populations. Using 30 microsatellites, a phylogeny was constructed using Shriver's distance (Dsw; Shriver *et al.*, 1995). This distance measure incorporates the SMM and is linear with respect to time. The results were consistent with previous studies based on classical nuclear markers and microsatellite loci (Nei and Roychoudhury, 1993; Bowcock *et al.*, 1994; DiRienzo *et al.*, 1994; Deka *et al.*, 1995), in showing greater divergence for African populations. However, the phylogeny based on microsatellite loci failed to show deep branches, indicating that the diversity among African populations is not as high as mitochondrial DNA analyses suggests (Jorde *et al.*, 1995).

Relationships between Human Populations Based on Microsatellite Data

Microsatellite loci can be used to infer the phylogenetic relationships among closely related taxa such as human populations due to the high rate of evolutionary change. Studies using microsatellite markers (Bowcock *et al.*, 1994; Deka *et al.*, 1995; DiRienzo *et al.*, 1994) reveal that the branching orders of human populations are very similar to studies that utilized classical nuclear DNA polymorphism's and mitochondrial DNA (Nei and Roychoudhury, 1993; Cann *et al.*, 1987). In addition, trees constructed using human individuals with microsatellite data reflect their geographic origin. This is in contrast to trees produced using mitochondrial DNA. The difference in the trees is thought to be due to the difference in the relative rates of evolutionary change for these markers (Bowcock *et al.*, 1994; Ruvolo, 1996).

Relationships Within Human Populations Based on Microsatellite Data

Microsatellite variation within populations have been studied in order to address questions regarding kinship, origins and the relationships between subpopulations.

Rower *et al.*, (1993) examined the microsatellite variation in a population of Yanomani Indians from Southern Venezuela. Significant differences were found in allele frequencies when compared to a control German population, reflecting the high degree of consanguineous mating and polygyny that are inherent in the social system of the Yanomani.

Lahermo *et al.*, (1996) examined the relationship between two Finno-Ugric speaking populations, the Finns and the Finnish-Saami (Lapps). These two populations occupy partly overlapping areas and may share a common history. However, cultural differences have separated the two populations for the last 1000 years. Furthermore, the origin of the Saami people is not known. Analysis of microsatellite loci and mitochondrial DNA revealed statistically significant differences between the genetic background of these two populations. Furthermore, genetic distance measures show that the Saami population is not closely related to their linguistic or geographic neighbors.

Objectives of the present study

The objectives of this thesis, are to determine whether microsatellite DNA can be used to detect any genetic differences between populations within India, and between India and other human populations.

These questions will be addressed by using microsatellite DNA. Microsatellite repeats are found primarily in non-coding regions of DNA and because of this, are thought to be selectively neutral. Along with high mutation rates and selective neutrality, microsatellites are believed to be good markers for studying the genetic relationships between closely related populations. Furthermore, genetic markers used in previous studies such as protein polymorphisms and mitochondrial DNA may not be selectively neutral. Six microsatellite markers will be used in this study to asses the genetic relationships between various populations.

Specifically, microsatellite loci will be used to detect any differences between North and South Indian populations which are divided by spoken language. Past studies using protein polymorphisms and more recently mitochondrial DNA have detected differences between the Indo-Aryan speaking populations of the North and the Dravidian speakers of the South.

Additionally, the microsatellite diversity of four regions within India - Northeast, Northcentral, Southeast and Southwest will be examined to determine if there is a correlation between genetic and geographic distance. Previous studies using protein polymorphisms have demonstrated that geographical distance may not necessarily be a determinant of genetic similarity or differentiation within India.

These microsatellite loci will also be used to asses the genetic relationships between the Indian population and other human populations. Previous studies have demonstrated that the Indian population is genetically closest to West Asian or Caucasian populations with some admixture from East Asian populations and is farthest away from African populations.

Finally, an evaluation of the use of six microsatellite markers in differentiating between populations will be made based on the results of this study.

Materials and Methods

Sampling Procedure

Blood samples were obtained from individuals of Indian origin residing in India or in Canada. Samples were obtained on a volunteer basis and each donor was asked to fill out a questionnaire in order to determine their place of birth, as well as their parent's place of birth.

Blood samples were collected in 5 ml VACUTAINER brand blood collection tubes containing 0.05 ml of 15% EDTA (K_3) solution. The blood was stored in 1 ml aliquots at -70°C.

Between 124 and 149 individuals were typed at six microsatellite loci for this study. The geographical origins of the samples based on the parents' birth state is shown in **figure 3**. The distribution of samples according to state and region is shown in **table 1**. Certain states were grouped into Northcentral, Northeast, Southeast and Southwest regions according to their proximity to each other as well as large sample availability.

Population data for the African, Caucasian and Hispanic groups used in this study were obtained from the Promega Technical Manual-TMD004 (1996). These populations were sampled from the United States. The Mexican and Asian population data was obtained from the United States aswell (Edwards *et al.*, 1992; Hammond *et al.*, 1994).

DNA Extraction

Genomic DNA was extracted from a 1 ml aliquot of whole blood. Red blood cells were lysed 3 times using an NH_4Cl/Na_4HCO_3 lysis buffer and the white blood pellet obtained was rinsed with a saline solution. The white blood pellet was lysed using SSTE and Proteinase K (Boehringer-Mannheim), followed by an incubation period for 1 hour at 55°C. Next, phenol:chloroform

Figure 3: Geographic distribution of samples.

Geographical location of 149 samples that were typed in this study. The exact location of each sample is unknown, beyond that of the state where their parents are from. The samples located within each state represent cases where the individuals parents both came from that state. The samples located on the border between two states represents individuals whose parents came from those states. The 8 samples located outside of the map represent individuals who were classified as North Indian. In this case, the parents came from Northern states that were not neighbors.

Sample Distribution

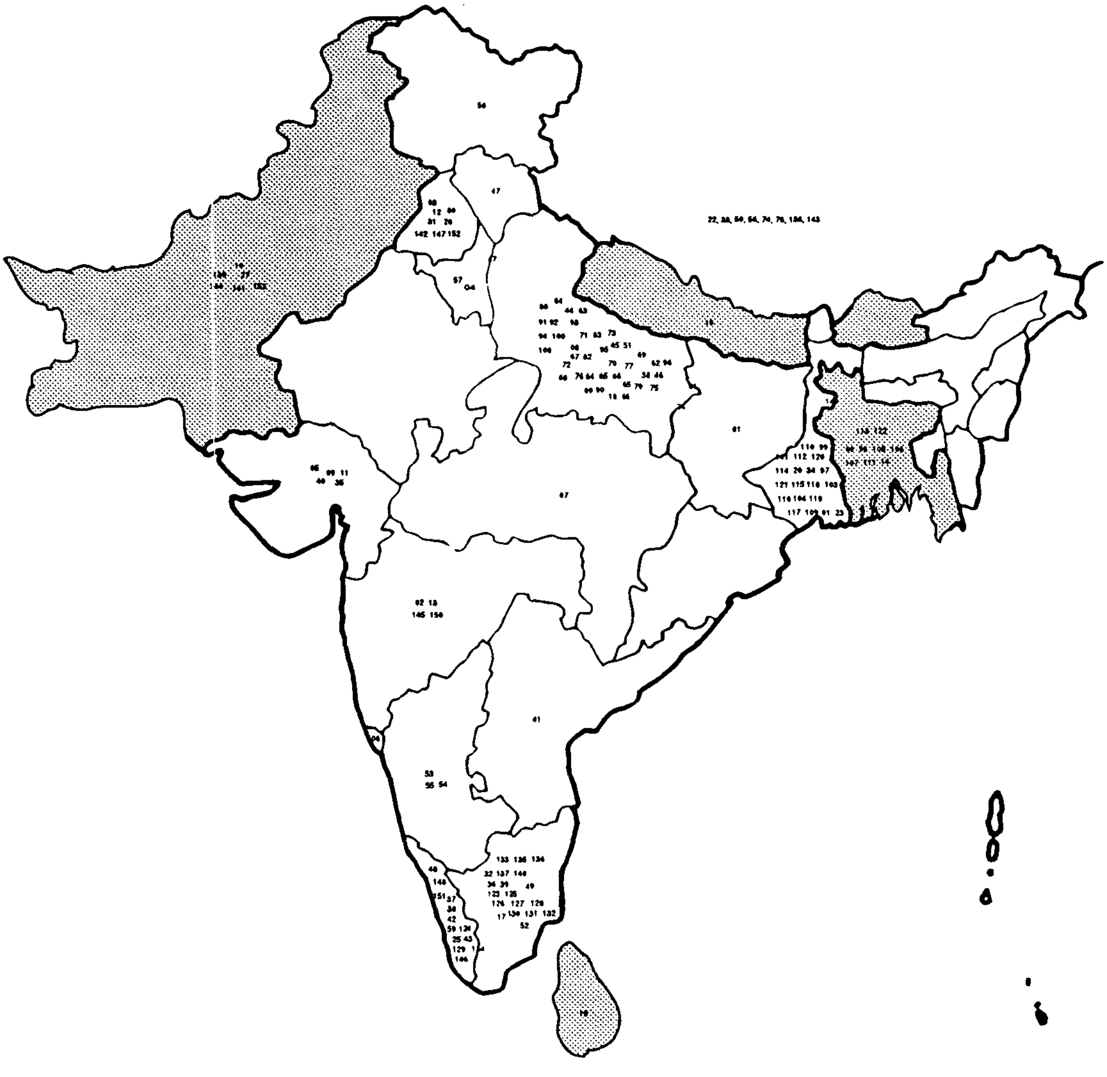


Table 1: Geographical Distribution of samples typed in this study.

NORTH INDIA	Sample Size	NORTHCENTRAL	NORTHEAST
Jammu + Kashmir	1		
Himachal Pradesh	1		
Punjab	8		
Haryana	2		
Uttar Pradesh	39	39	
Bihar	1	1	
Nepal	1	1	
West Bengal	21		21
Bangladesh	9		9
Pakistan	5		
Gujarat	5		
Maharashtra	4		
Goa	1		
Madhya Pradesh	1		
Neighbors	5	2	1
Northern	8		
Total	**112**	**43**	**31**
SOUTH INDIA		**SOUTHEAST**	**SOUTHWEST**
Sri Lanka	1	1	
Andhra Pradesh	1		
Karnataka	3		3
Tamil Nadu	19	19	
Kerala	12		12
Neighbors	1		1
Total	**37**	**20**	**16**
Grand Total	**149**		

extractions consisting of two extractions using equal volumes of equilibrated phenol, one extraction using a 1:1 mixture of phenol:chloroform iso-amyl alcohol, and one extraction using only chloroform iso-amyl alcohol was performed. The DNA was then precipitated with 2-3 volumes of 100% ethanol and 1/10th volume of 2 M NaAc. The DNA pellets were rinsed with 70% ethanol, dried and dissolved in 50 ul of TE. The concentrated DNA samples were stored at -20°C. The DNA samples were quantified using a Flurometer (Hoefer Scientific Instruments: Model TKO100). Aliquots of the concentrated DNA samples were diluted to approximately 20 ng/ul using double distilled H_2O (DDW). The diluted samples were used in the Polymerase Chain Reaction.

PCR Amplification of 6 Microsatellite Loci

The Polymerase Chain Reaction was used to amplify 6 microsatellite loci: CSF1PO, TPOX, TH01, FESFPS, F13A01 and vWA. They are located on different chromosomes and consist of tetramer repeats. **Table 2** gives the name, chromosomal location, repeat sequence and the number of alleles found at each locus.

PCR amplifications were carried out in thin-walled GeneAmp™ (Perkin Elmer) reaction tubes using 2 GenePrint™ STR Multiplex kits (Promega). Each of these kits contained 10X Primer Pairs as a mixture, for the simultaneous amplification of 3 microsatellite loci in one reaction tube. Due to patentcy regulations, the sequence of the primer pairs were not revealed. In addition to the primers, each of the kits contained a mixture of the allelic ladders for the same set of 3 loci, 10X buffer mix, positive control DNA (K562 High Molecular Weight), 2X loading solution and pGEM® DNA markers.

Reactions were carried out in 25 ul volumes following the manufacturer's instructions. Each reaction contained 17.35 ul of sterile water, 2.5 ul of 10X buffer (500 mM KCl, 100 mM Tris-HCl, pH 9.0 at 25°C, 15 mM $MgCl_2$, 1% Triton® X-100, 2 mM of each dNTP), 2.5 ul of 10X primer pairs (5 uM), 0.15 ul of Taq DNA Polymerase (0.75 Units) (Boehringer Mannheim) and 2.5 ul (approximately 50 ng) of template DNA. A negative and positive control (25 ng K562 DNA) was included.

Table 2: Locus-specific information of the 6 microsatellite loci used in this study.

STR Locus	Chromosomal Location	Repeat Sequence	Allele Repeats
CSF1PO	5q33.3-34	AGAT	7 - 15
TPOX	2p23-2pter	AATG	6 - 13
TH01	11p15.5	AATG	5 -11
F13A01	6p24-25	AAAG	4 - 16
FESFPS	15q25-qter	AAAT	7 - 14
vWA	12p12-pter	AGAT	13 - 20

Reactions were overlaid with one drop of mineral oil and briefly centrifuged to bring the contents to the bottom of the tube.

The cycling reactions were performed using the Perkin Elmer Cetus DNA Thermal Cycler 480. The GenePrint™ STR Multiplex kit containing the loci, CSF1PO, TPOX and TH01 used the following thermal cycling parameters: Denaturation at 96°C for 2 minutes followed by 10 of denaturing, annealing and extension cycles of 1 minute at 94°C, 1 minute at 64°C and 1.5 minutes at 70°C. Following this, an additional 20 denaturing, annealing and extension cycles of 1 minute at 90°C, 1 minute at 64°C and 1.5 minutes at 70°C was performed. The second GenePrint™ STR Multiplex kit containing the loci, FESFPS, F13A01 and vWA used the following cycling protocol: 96°C for 2 minutes, then 94°C for 1 minute, 60°C for 1 minute, 70°C for 1.5 minutes for 10 cycles, then: 90°C for 1 minute, 60°C for 1 minute, 70°C for 1.5 minutes for 20 cycles, then: 60°C for 30 minutes. **Table 3** shows the PCR product sizes for each locus, alleles at each locus and the positive control's genotype (Promega Technical Manual-TMD004, 1996).

PCR products were visualized by running 5 ul of each reaction mixed with 2 ul of 5X loading solution run on a 2% agarose minigel, stained with ethidium bromide and illuminated by ultraviolet light. DNA fragments were sized in comparison to a 100 bp ladder (GIBCO-BRL). **Figure 4** shows a picture of the agarose gel where 3 amplified microsatellite loci are visible.

Polyacrylamide Gel Electrophoresis

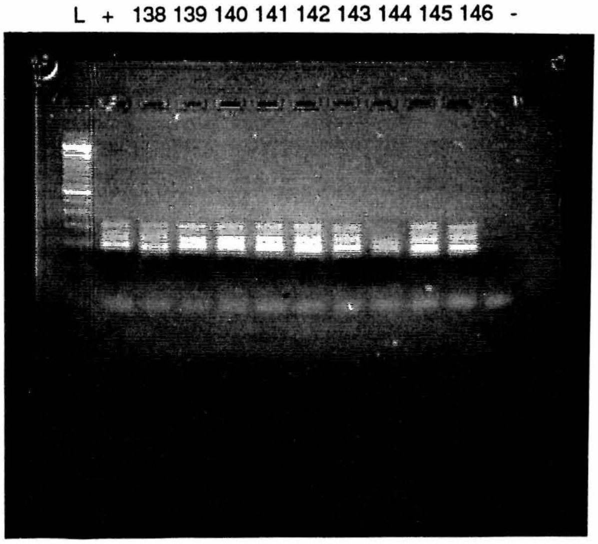
The PCR products were run on 6% polyacrylamide gels containing 7 M urea using the reagents contained in the GenePrint™ STR Multiplex kits. The gels were run on the Life Technologies Inc. Model S2 sequencing apparatus at 2000 Volts, 65 Watts and 65 mAmps. PCR products were prepared by mixing 2.5 ul of DNA sample with 2.5 ul of 2X loading solution. In addition to the DNA samples, 2.5 ul of positive control, ladder mix and pGEM® were mixed with 2.5 ul of 2X loading

Table 3: PCR product size and genotypes of the of the positve control at 6 microsatellite loci used in this study.

STR Locus	PCR Product	K-562 Genotype
CSF1PO	295-327 bp	10 / 9
TPOX	224-252 bp	9 / 8
TH01	179-203 bp	9.3 / 9.3
F13A01	283-331 bp	5 / 4
FESFPS	222-250 bp	12 / 10
vWA	139-167 bp	16 / 16

Figure 4: Photograph of agarose gel.

Photograph shows the products of a multiplex PCR reaction. Samples 138 to146, the 100 bp ladder as well as the positive and negative control can be seen. The three bands from the top to bottom correspond to the loci CSF1PO, TPOX and TH01 respectively. The 2% agarose gel containing EtBr was run in TAE buffer and visualized under a UV transilluminator (302 nm).


L + 138 139 140 141 142 143 144 145 146 -

solution. The samples were spun briefly in a microcentrifuge and denatured at 95°C for 2 minutes immediately before loading the samples onto the gel. The gel was run for approximately 1.5 hours (Promega Technical Manual-TMD004, 1996).

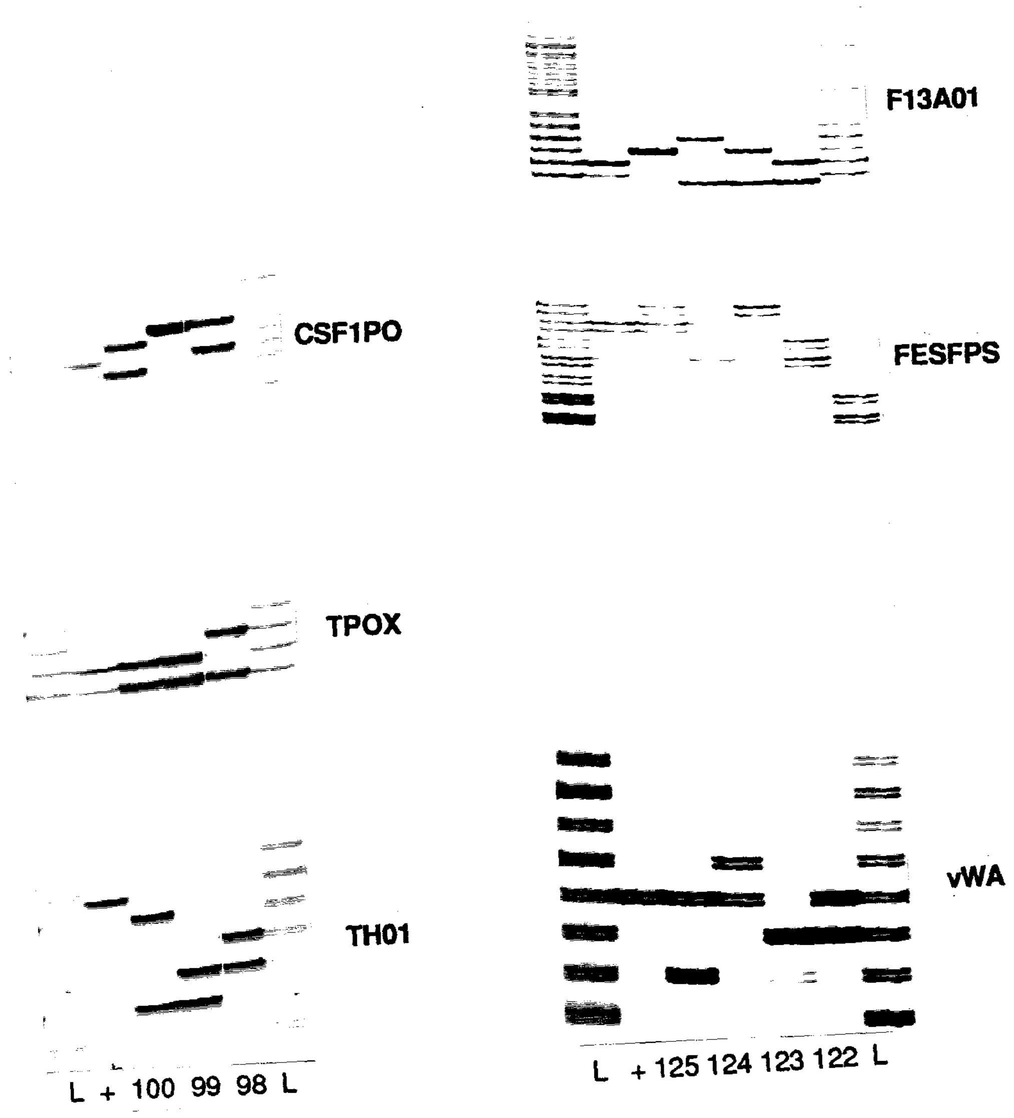
Silver Staining Polyacrylamide Gels

Following the polyacrylamide gel electrophoresis, the glass plates were separated, with the gel adhering to the short glass plate. Gels were stained using the DNA Silver Staining System (Promega) kits with a few modifications of the manufacturer's instructions.

All the steps of the staining procedure required the plate containing the gel to be agitated gently. Immediately following electrophoresis, the gel was placed in 2000 ml of a fix/stop solution (200 ml glacial acetic acid, 1800 ml deionized H_2O) for 20 minutes, followed by 3 rinses with approximately 1300 ml of deionized H_2O for 2 minutes each. Next, the gel was placed in 2000 ml of a staining solution (2 g silver nitrate, 3 ml 37% formaldehyde, 2000 ml deionized H_2O) for 30 minutes, then briefly dipped in 1300 ml of deionized H_2O for 5-10 seconds. The gel was then placed in 1000 ml of a developer solution (60 g sodium carbonate, 3 ml 37% formaldehyde, 400 ul sodium thiosulphate, 2000 ml deionized H_2O) chilled to between 4-10°C for approximately 5 minutes or until the alleles and ladders were visible. Next, the gel was placed in 1000 ml of fresh developer solution until the alleles became more visible at which time, 1000 ml of fix/stop solution was added directly to the developer solution in order to stop the developing reaction for 5 minutes. Finally, the gel was rinsed twice with 1300 ml of deionized H_2O for 1 minute each. The gel was left standing upright to dry overnight. A picture of the gel was produced using Automatic Processor Compatible Film (Promega) which gives a mirror image of the gel. The alleles were scored directly from the dried gel or from the photograph. **Figure 5** shows the alleles or genotypes of the samples at the 6 microsatellite loci, along with the allelic ladders and the positive control (Promega Technical Manual-TM023, 1996).

Figure 5: Photographs of polyacrylamide gels.

Photographs show the products of two separate multiplex PCR reactions. The alleles or genotypes of each sample and the positive control can be seen. For allelic ladder designations and positive control genotypes, see **tables 2** and **3**. The first multiplex PCR reaction contained the loci CSF1PO, TPOX and TH01 and the second reaction contained the loci F13A01, FESFPS and vWA. The 6% polyacrylamide gels were run in TBE buffer and visualized by silver staining.


F13A01
FESFPS
CSF1PO
TPOX
TH01
vWA
L + 100 99 98 L
L + 125 124 123 122 L

See Appendix for detailed protocols of the procedures described in this section and allele frequencies and of each locus and the genotypes of the samples used in this study.

Heterozygosity

Observed and unbiased estimates of expected heterozygosity were calculated for the Indian, African, Caucasian and Hispanic populations at six microsatellite loci and for the Mexican and Asian populations at four microsatellite loci based on genotypic and allele frequency data. The source and sample size of this data is shown in **Table 4**. Observed heterozygosity was caluculated as the number of observed heterozygote genotypes in the population divided by the total number of individuals in the population. Unbiased estimates of expected heterozygosities were calculated using the following formula:

$$He = [\, n(1 - \sum_{i=1}^{k} p_i^2)\,] / n - 1$$

where n is the number of genes sampled and p_i is the frequency of the ith allele (Nei, 1978).

Hardy-Weinberg Equilibrium

Three statistical tests were used to determine whether there was any discordance of genotypic frequencies from their Hardy-Weinberg expectations for the Indian population at six microsatellite loci.

The chi-square Goodness-of-fit test was used.

$$\chi^2 = \Sigma\,(\,O - E\,)^2 / E$$

Genotypic classes, where the expected counts were less than 5 were lumped together with adjacent classes to give a minimum expected value of 5 or greater. The number of degrees of freedom was calculated as the number of classes - 1.

Table 4: Source and sample sizes of populations used in this study.

	Locus					
Population	**CSF1PO**	**TPOX**	**TH01**	**F13A01**	**FESFPS**	**vWA**
Indian	148	148	148	125	124	140
N. Indian	*111*	*111*	*111*	*95*	*94*	*103*
S. Indian	*37*	*37*	*37*	*30*	*30*	*37*
African[1]	202	204	204	218	217	218
Caucasian[1]	209	209	209	209	212	212
Hispanic[1]	216	216	216	222	210	211
Mexican[2,3]	187	-	192	183	159	-
Asian[2,3]	72	-	77	63	67	-

1. Promega Technical Manual-TMD004, 1996
2. Edwards *et al*., 1992
3. Hammond *et al*., 1994

The second Goodness-of-fit test used was the log-likelihood ratio test, also known as the G-test. The formula for the G-test is:

$$G = 2[\overset{a}{\Sigma} f_i \ln f_i - \overset{a}{\Sigma} f_i \ln \hat{f}_i]$$

where f_i is the observed count, $\hat{f}_i$ is the expected count under Hardy-Weinberg expectations and a is the number of classes (Sokal and Rohlf, 1969). Genotypic classes that had expected counts less than 5 were pooled with adjacent classes to give a minimum expected value of 5 or more. The number of degrees of freedom was calculated as the number of classes - 1. The G-test statistic appears to follow the χ^2 distribution a bit more closely than the X^2 test statistic when sample sizes are small (Sokal and Rohlf, 1969).

The third statistical test was performed using the program CHIHW (Zaykin and Pudovkin, 1991). This program estimates the probability of the null-hypothesis (agreement with Hardy-Weinberg Equilibrium) using a Monte-Carlo simulation. Lumping data to increase expected counts can potentially lead to a decrease in the level of significance for a particular data set. Therefore, the solution to lumping or loss of information, is to generate using a Monte-Carlo simulation a distribution of χ^2 expected if the null hypothesis were true for the particular data set under study (Roff and Bentzen, 1989). This is done by generating random samples with the same numbers of genes of each category as in the original sample under analysis. For each generated table, χ^2 is computed and compared with χ^2 for the original table. The result of the Monte-Carlo simulation, is the ratio of generated tables for which χ^2 is greater than or equal to χ^2 original, to the total number of runs, in this case, 1000. This ratio is regarded as the probability (P) of the null hypothesis for the original sample (Zaykin and Pudovkin, 1991). The significance of all three tests was determined at the 5% level.

Contingency Analysis

Homogeneity tests for pairwise contingency tables of allele frequency distributions were performed. Three types of contingency analyses were done. The first compared the Indian population in a pairwise fashion to the African, Caucasian, Hispanic, Mexican and Asian populations at four (CSF1PO, TH01, F13A01, FESFPS) and six microsatellite loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA). The second comparison was between the North and South Indian populations at six microsatellite loci and thirdly, pairwise comparisons were done between the four regions within India (Northcentral, Northeast, Southwest, Southeast) using six microsatellite loci. **Figures 6** and **7** shows the North-South division and the four regions of the Indian subcontinent respectively, along with the sample sizes used.

Three statistical tests were used to test for homogeneity (null hypothesis) between pairs of populations.

The chi-square Goodness-of-fit test was calculated for each cell as follows:

$$\chi^2 = (O - E)^2 / E$$

where E(expected count) = row total x column total / grand sum and χ^2 is the grand sum of χ^2 values for each cell. The number of degrees of freedom was calculated as (# rows - 1)(# columns - 1) (Sokal and Rohlf, 1969).

The second statistical method used, was the G-test and was calculated as follows:

$$G = 2[\ (\Sigma f_i \ln f_i \text{ cell counts}) - (\Sigma f_i \ln f_i \text{ row and column totals}) + n\ln n\ ]$$

where f_i is the observed count and n is the grand total. The number of degrees of freedom was calculated as it was for the χ^2 test. For both the χ^2 and G-test, if the expected counts were less than 5, adjacent cells were pooled until a minimum expected value of 5 or more was reached and observed counts were pooled to match the expected (Sokal and Rohlf, 1969).

The third method of analysis utilized the program CHIRXC (Zaykin and Pudovkin, 1991) whereby, a Monte-Carlo simulation was performed. This program generates random RxC

Figure 6: Northern and Southern geographical regions of India.

Geographic division of the Indian subcontinent based on the two major language families. Northern regions of the Indian subcontinent speak languages of Indo-Aryan origin while Southern regions speak languages of Dravidian origin.

Geographical Regions

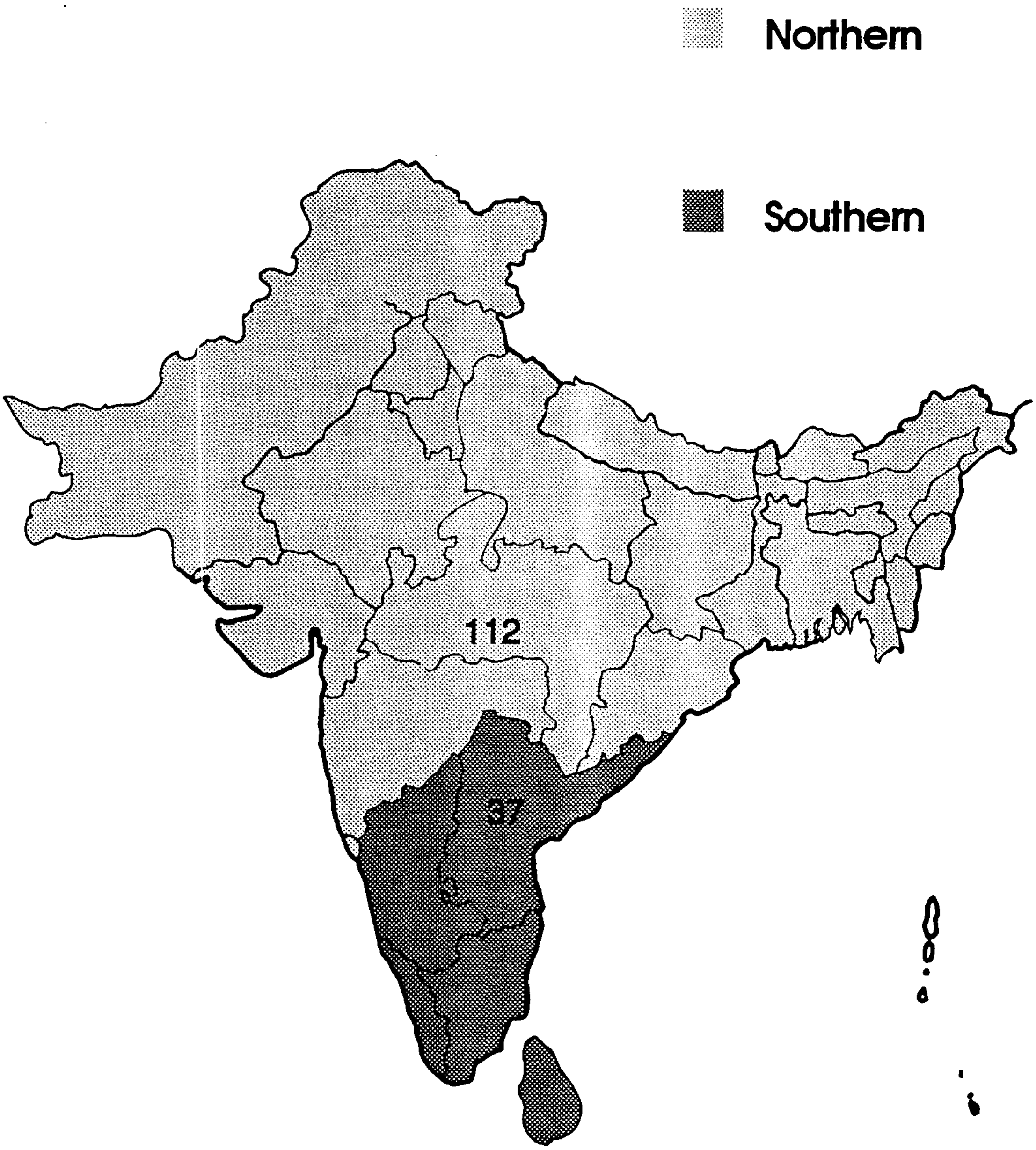
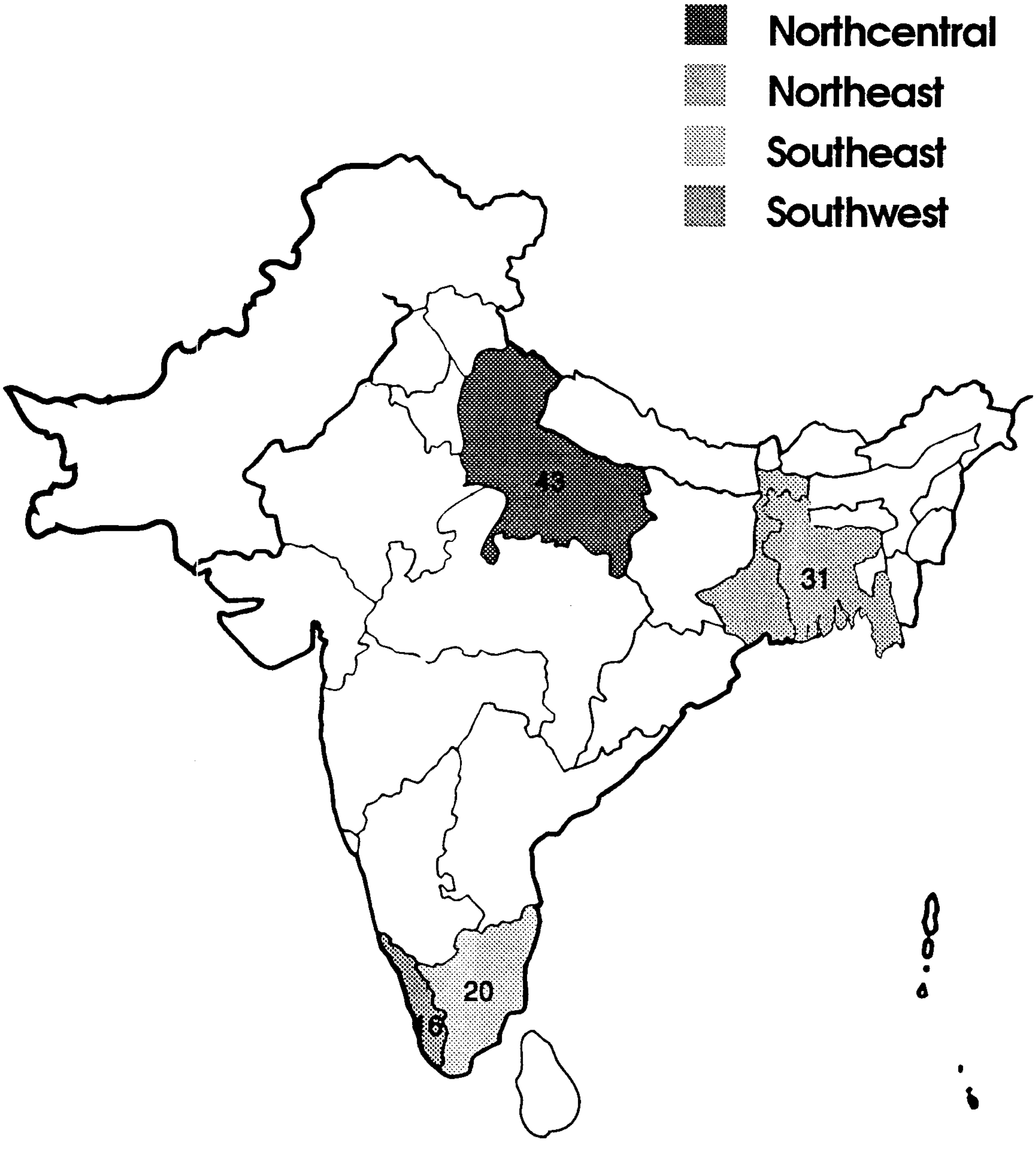


Figure 7: Geographical subdivisions of India.

Geographic subdivisions of India were based on the regions where a greater number of samples were found. In order to increase the sample sizes of these regions, a few samples from neighboring states were pooled with the larger group.

Geographical Regions



contingency tables with the same marginal totals as in the original RxC table under analysis. For each generated table, χ^2 was computed and compared with the χ^2 value for the original table. The result of the Monte-Carlo simulation, is the ratio of numbers of generated tables for which χ^2 was greater than or equal to the original χ^2, to the total number of generated tables or number of runs, in this case, 1000. This ratio is regarded as the probability (P) of the null hypothesis for the original data set (Zaykin and Pudovkin, 1991). The significance of all three tests was determined at the 5% level.

Descriptive Statistics

The mean allele size and standard deviation, mode, median, minimum and maximum allele, was computed for allele frequency distributions for the Indian (North and South), African, Caucasian, Hispanic, Mexican and Asian populations at each microsatellite locus. The purpose of the descriptive statistics was to note any trends in the data and to potentially infer a correlation with the results of the contingency analysis.

Genetic Distance

Genetic distance was computed using Goldstein's *et al.*, (1995) measure $(\delta\mu)^2$, between pairs of populations within India(Northcentral, Northeast, Southwest, Southeast) and between other populations(Indian, African, Caucasian, Hispanic, Mexican, Asian) using four (CSF1PO, TH01, F13A01, FESFPS) and six (CSF1PO, TH01, TPOX, F13A01, FESFPS, vWA) microsatellite loci. The formula for genetic distance is:

$$(\delta\mu)^2 = (\mu A - \mu B)^2$$

where μA is the mean allele size in population A and μB is the mean allele size in population B. This distance for microsatellite loci is based on the Stepwise Mutation Model. Genetic distance is calculated between pairs of populations at each locus and then is averaged over all loci (Goldstein *et* *al.*, 1995). Distance measures were computed using the program MICROSAT 1.4 (Minch *et al.*, 1994). Standard errors of distance measures were obtained by bootstrapping the data by locus 1000 times.

Phylogenetic Analysis

The phylogenetic relationship between populations was examined using four (CSF1PO, TH01, F13A01, FESFPS) and six (CSF1PO, TH01, F13A01, FESFPS, vWA) microsatellite loci. Using six loci, the relationships between the Indian (North and South), African, Caucasian and Hispanic populations was examined and using four loci, two additional populations, Mexican and Asian were added to the previous four groups. Goldstein's $(\delta\mu)^2$ genetic distance measure was computed using the program MICROSAT 1.4 (Minch, *et al.*, 1994). The distance matrices produced in this program were used with the PHYLIP software package (Felsenstein, 1993). All distance trees produced were unrooted and were inferred using the Neighbor-Joining method (Saitou and Nei, 1987) using the program Neighbor. Consensus trees were obtained by bootstrapping the data in MICROSAT 1.4, 1000 times, then, utilizing the programs Neighbor and CONSENSE. Consensus trees gave bootstrap or significance values. The distance trees were rearranged using RETREE and drawn using TREETOOL 1.0 (Macrukanas, 1991) and the resulting postscript files were edited to include the bootstrap values and labels.

Correlation

A correlation between genetic distance $(\delta\mu)^2$ and geographic distance (km) was conducted between four regions within India (Northcentral, Northeast, Southwest, Southeast). The formula used was Pearson's product moment correlation coefficient:

$$r = Sxy / SxSy$$

where Sx and Sy are the standard deviations of x (geographic distance) and y (genetic distance) respectively. The correlation coefficient is a value between +1, indicating a perfect positive correlation, and -1, indicating a perfect negative correlation. The significance of r was determined. It is determined by testing whether the sample correlation coefficient r could have come from a population with a parametric correlation coefficient of zero. Therefore, the null hypothesis is $H_o : \rho = 0$. This means that the two variables are not correlated. If the absolute value of the observed r is greater than the tabulated value of the critical value of r, the null hypothesis can be rejected. The number of degrees of freedom was calculated as the number of samples - 2 (Sokal and Rohlf, 1969).

Fst

Fst values were computed by pooling four populations (Indian, African, Caucasian, Hispanic) at six microsatellite loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA) and by pooling six populations (Indian, African, Caucasian, Hispanic, Mexican, Asian) at four microsatellite loci (CSF1PO, TH01, F13A01, FESFPS). Reynolds *et al.*, (1983) formula of $Fst = -\ln (1 - \theta)$ was used. Theta(θ) is the coancestry coefficient and distance measures such as Fst, are designed to measure the divergence between populations that is caused by drift. For this reason, Fst is considered to be an appropriate measure for short-term evolution when mutation can be neglected, and for this reason also, Fst is expected to be better in smaller populations (Reynolds *et al.*, 1983). Fst measures based on θ is a variance based method for estimating Fst. In Reynolds *et al.*, (1983) formula,

$$\theta = a / a + b$$

where a is the between subpopulation variance in allele frequency and b is the between individuals within subpopulation variance in allele frequency. Fst values were computed using the program MICROSAT 1.4 (Minch *et al.*, 1994).

Results

Heterozygosity

Observed and expected heterozygosities were calculated for the Indian, African, Caucasian and Hispanic populations at six microsatellite loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA), and at four microsatellite loci (CSF1PO, TH01, F13A01, FESFPS) for the Mexican and Asian populations. **Table 5** shows the heterozygosity values and sample sizes for each population at each locus. Observed heterozygosity values across all six loci were between 0.61 and 0.83. The observed and expected heterozygosities for each population at each locus were close. The average observed heterozygosity over six loci was the highest in the African population (H = 0.79), followed by the Caucasian, Indian and Hispanic populations.

Hardy-Weinberg Equilibrium

Deviations of genotypic frequencies from Hardy-Weinberg expectations were tested for in the Indian population, using three different statistical tests. The results of these tests are shown in **table 6**. The Monte-Carlo test revealed significant deviations ($P < 0.05$) from Hardy-Weinberg expectations at the TPOX ($P = 0.024$) and vWA ($P = 0.003$) loci. Chi-square and G-test results were not significant at these loci. The other four loci did not reveal any significant deviations from Hardy-Weinberg expectations.

Table 5: Observed and Expected Heterozygosity values and Sample Sizes of Populations.

	Population					
Locus	**Indian**	**African[1]**	**Caucasian[1]**	**Hispanic[1]**	**Mexican[2,3]**	**Asian[2,3]**
CSF1PO						
n	148	202	209	216	187	72
Ho	0.73	0.83	0.79	0.69	-	-
He	0.73	0.79	0.73	0.72	0.73	0.75
TPOX						
n	148	204	209	216	-	-
Ho	0.70	0.74	0.66	0.68	-	-
He	0.69	0.78	0.64	0.67	-	-
TH01						
n	148	204	209	216	192	77
Ho	0.82	0.73	0.77	0.75	0.70	0.61
He	0.79	0.76	0.78	0.77	0.76	0.72
F13A01						
n	125	218	209	222	183	63
Ho	0.76	0.78	0.77	0.80	-	-
He	0.78	0.81	0.75	0.80	0.79	0.65
FESFPS						
n	124	217	212	210	159	67
Ho	0.69	0.80	0.70	0.68	-	-
He	0.71	0.76	0.68	0.70	0.70	0.70
vWA						
n	140	218	212	211	-	-
Ho	0.78	0.83	0.83	0.75	-	-
He	0.80	0.81	0.81	0.79	-	-
Avg. Ho	0.75	0.79	0.74	0.73	-	-

[1] Promega Technical Manual-TMD004, 1996

[2] Edwards *et al.*, 1992

[3] Hammond *et al.*, 1994

Table 6: Test of Hardy-Weinberg Equilibrium for the Indian Population at Six Microsatellite Loci.

	Statistic					
Locus	**X^2 - Test**		**G - Test**		**M.C. - Test**	
CSF1PO	X^2	= 1.10	G	= 0.58	P	= 0.474
	df	= 9	df	= 9		
	P	> 0.05	P	> 0.05		
TPOX	X^2	= 1.42	G	= 1.20	*P	= 0.024
	df	= 7	df	= 7		
	P	> 0.05	P	> 0.05		
TH01	X^2	= 6.11	G	= 6.00	P	= 0.628
	df	= 11	df	= 11		
	P	> 0.05	P	> 0.05		
F13A01	X^2	= 6.09	G	= 6.00	P	= 0.095
	df	= 9	df	= 9		
	P	> 0.05	P	> 0.05		
FESFPS	X^2	= 2.70	G	= 2.60	P	= 0.519
	df	= 7	df	= 7		
	P	> 0.05	P	> 0.05		
vWA	X^2	= 4.29	G	= 4.20	*P	= 0.003
	df	= 14	df	= 14		
	P	> 0.05	P	> 0.05		

* Significant $P < 0.05$.

Contingency Analysis and Descriptive Statistics

Relationship between North and South India

Homogeneity tests for pairwise contingency tables of allele frequency distributions were performed between North and South India at six microsatellite loci. The results of the contingency analysis are shown in **table 7** and histograms comparing the allele frequency distributions of North and South India at six loci are shown in **figures 8 - 13**. The Monte-Carlo simulation revealed a significant difference ($P < 0.05$) in the allele frequency distribution between the North and South Indian populations at the vWA locus with $P = 0.033$. However, analysis using the χ^2 and G-test were not significant at this locus. The other five loci did not show any significant deviations from homogeneity of the allele frequency distributions between these two populations.

Descriptive statistics were computed for the allele frequency distributions in order to detect any similarities and differences between populations, or to detect any features of the distribution(s) that was unique to a particular population. Some results of the descriptive statistics were used to infer a correlation with the results of the contingency analysis. However, those features of a distribution that are considered similar, different or unique to a particular population(s) may or may not be the cause of a significant or non-significant result in the contingency analysis. Rather, it may be one, or a combination of features that yields a significant or non-significant result. Therefore, the results of the descriptive statistics are used to visualize any trends in the data and may not necessarily suggest a causal relationship with the contingency analysis.

The descriptive statistics of the North and South Indian populations are shown in **table 8**. Generally, there was little difference in the mean allele size of each locus between the North and South Indian populations. However, the mode at four (CSF1PO, TPOX, TH01, vWA) out of six loci was different between the two populations. The total number of alleles found at each locus

Table 7: Contingency Table Analysis of Allele Frequency Distributions for North and South India at Six Microsatellite Loci.

	Statistic					
Locus	**X^2 - Test**		**G - Test**		**M.C. - Test**	
CSF1PO	X^2	= 6.50	G	= 6.40	*P*	= 0.361
	df	= 3	*df*	= 3		
	P	> 0.05	*P*	> 0.05		
TPOX	X^2	= 4.73	G	= 4.60	*P*	= 0.188
	df	= 3	*df*	= 3		
	P	> 0.05	*P*	> 0.05		
TH01	X^2	= 1.95	G	= 1.80	*P*	= 0.751
	df	= 4	*df*	= 4		
	P	> 0.05	*P*	> 0.05		
F13A01	X^2	= 5.32	G	= 5.6	*P*	= 0.201
	df	= 4	*df*	= 4		
	P	> 0.05	*P*	> 0.05		
FESFPS	X^2	= 0.93	G	= 0.40	*P*	= 0.839
	df	= 3	*df*	= 3		
	P	> 0.05	*P*	> 0.05		
vWA	X^2	= 7.98	G	= 8.2	**P*	= 0.033
	df	= 5	*df*	= 5		
	P	> 0.05	*P*	> 0.05		

* Significant $P < 0.05$.

Figure 8: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the North and South Indian populations at the CSF1PO locus.

Allele Frequency Distribution at the CSF1PO Locus for North and South India

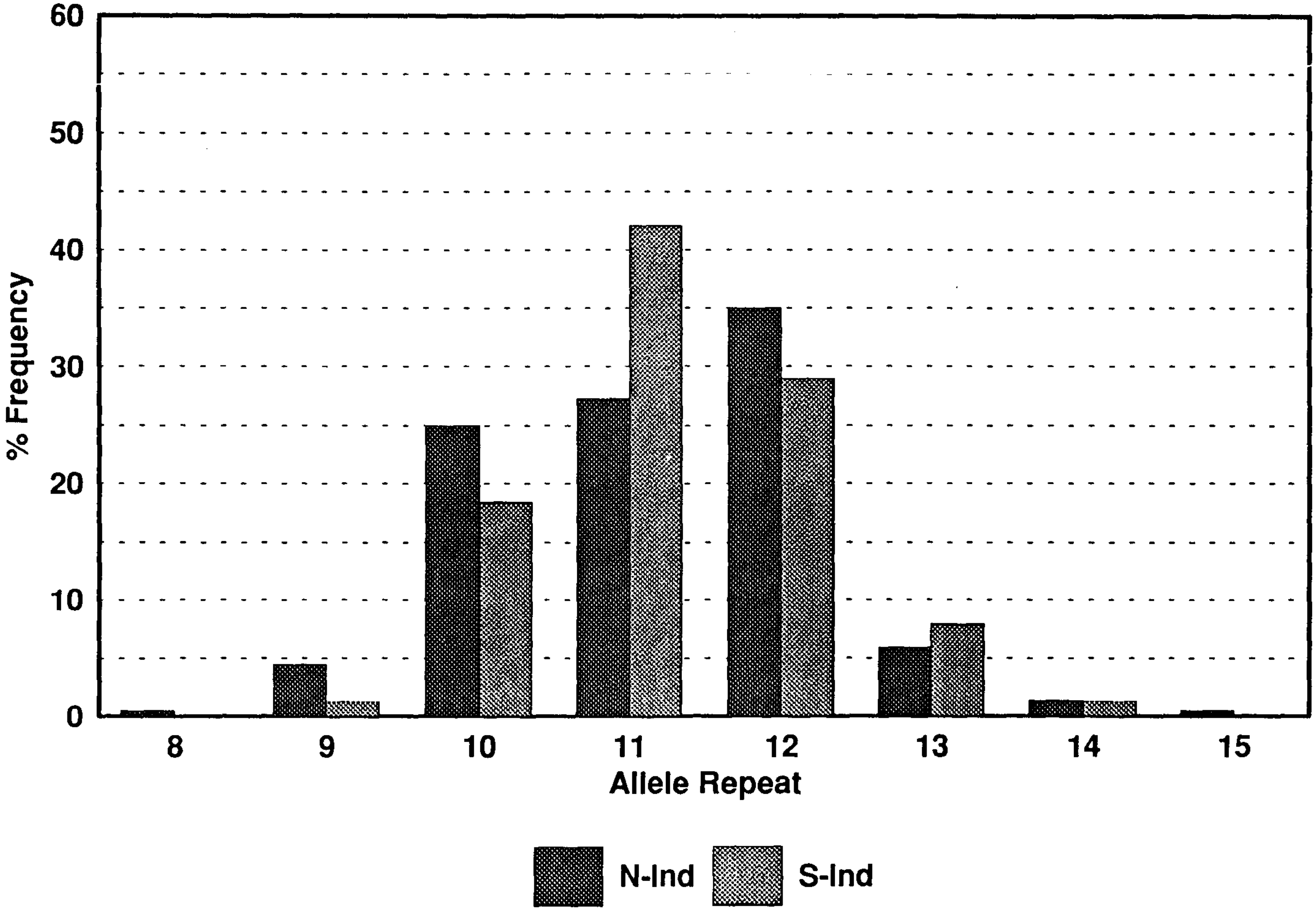


Figure 9: Histograrn of allele frequency distribution.

Histograrn of allele frequency distributions of the North and South Indian populations at the TPOX locus.

Allele Frequency Distribution at the
TPOX Locus for North and South India

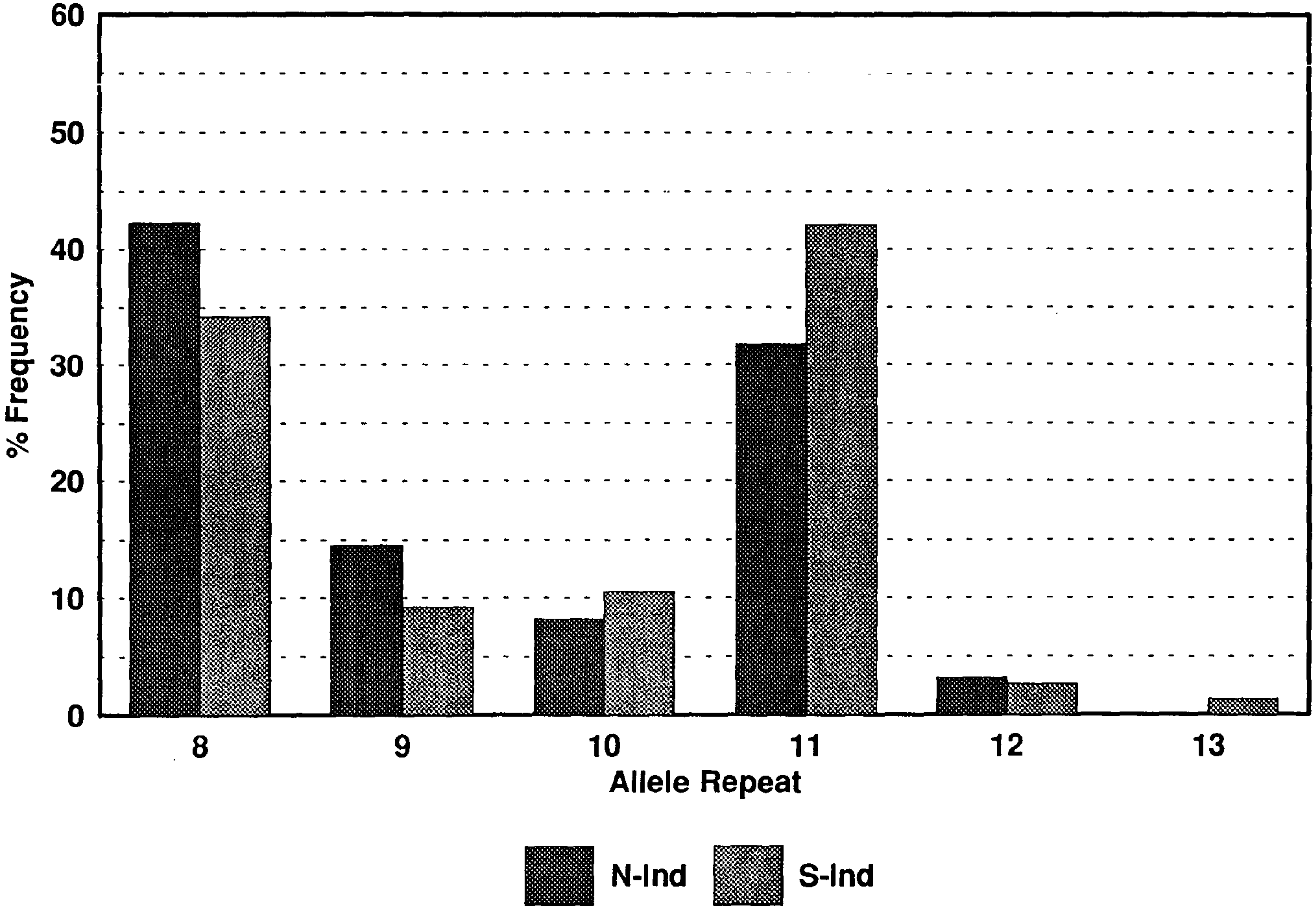


Figure 10: Histograrn of allele frequency distribution.

Histograrn of allele frequency distributions of the North and South Indian populations at the TH01 locus.

Allele Frequency Distribution at the
TH01 Locus for North and South India

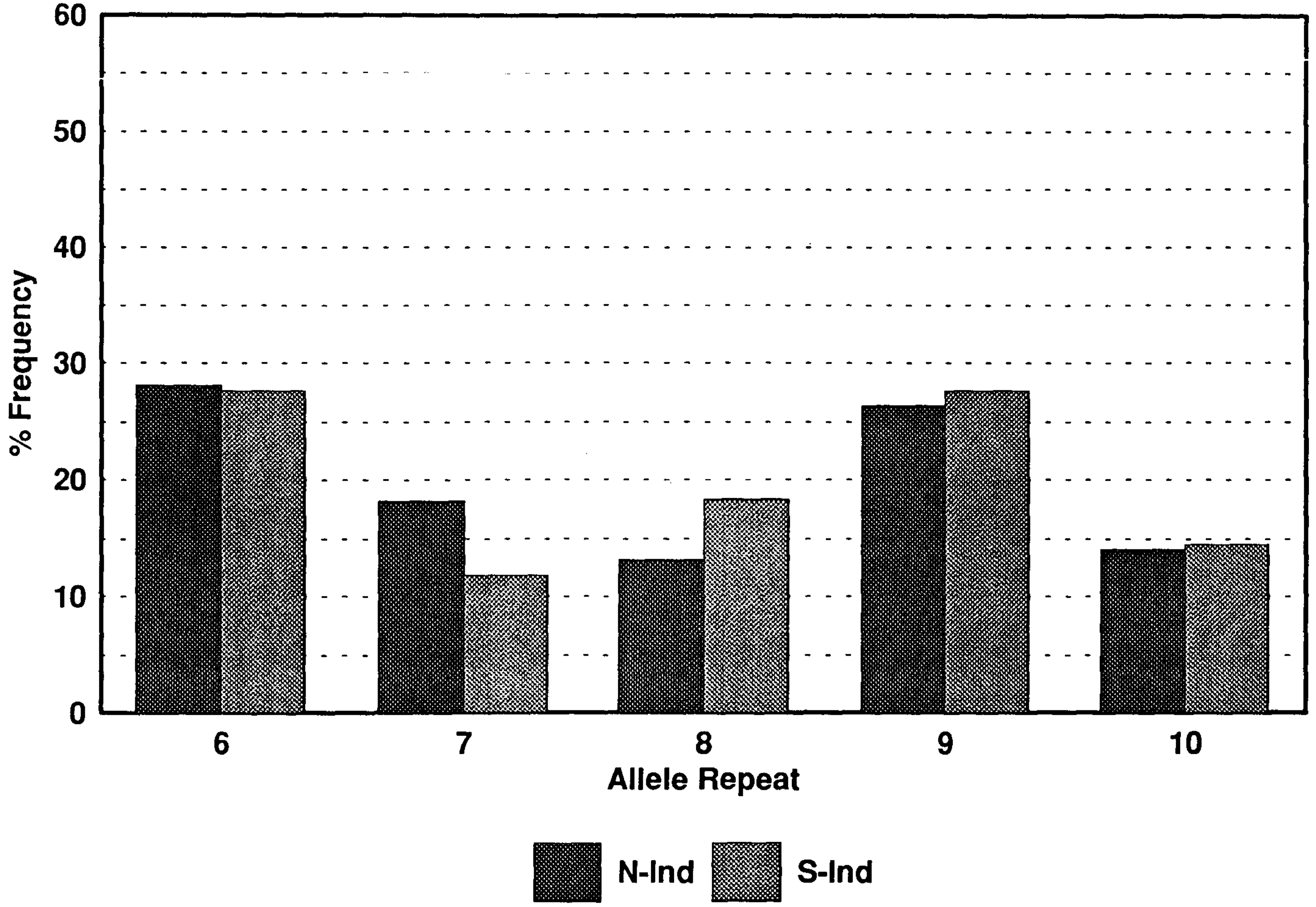


Figure 11: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the North andSouth Indian populations at the F13A01 locus.

Allele Frequency Distribution at the F13A01 Locus for North and South India

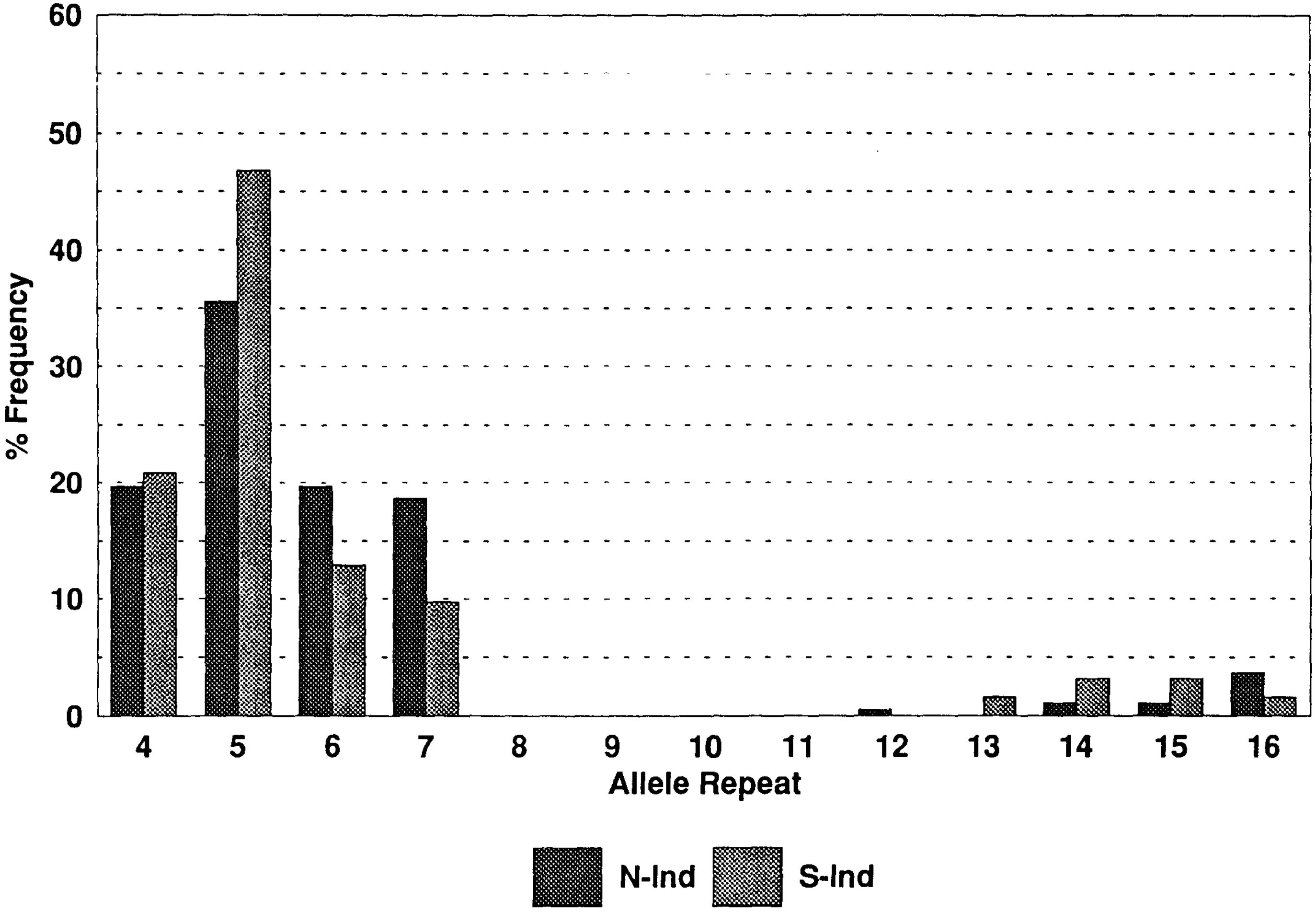


Figure 12: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the North andSouth Indian populations at the FESFPS locus.

Allele Frequency Distribution at the
FESFPS Locus for North and South India

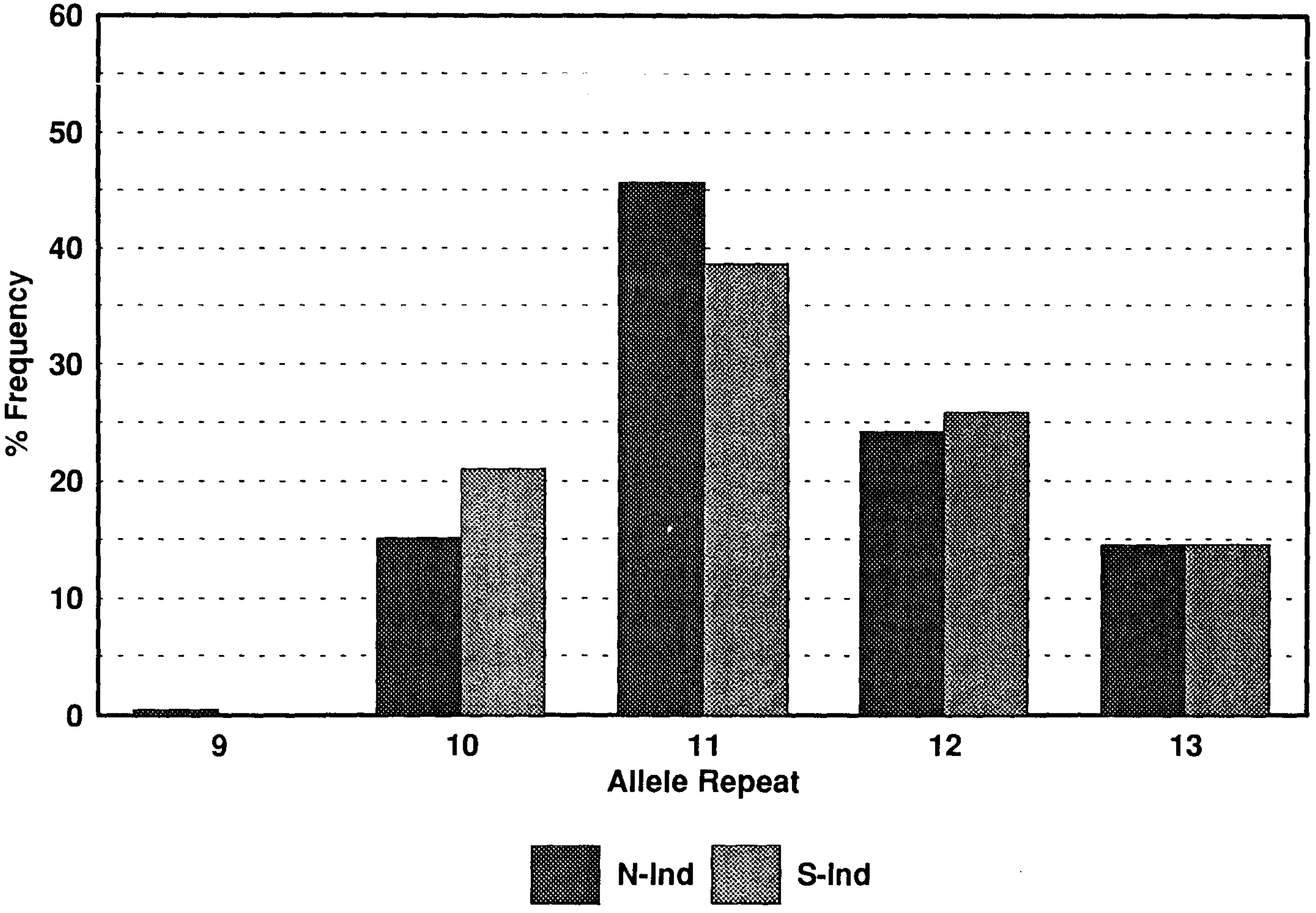


Figure 13: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the North and South Indian populations at the vWA locus.

Allele Frequency Distribution at the
vWA Locus for North and South India

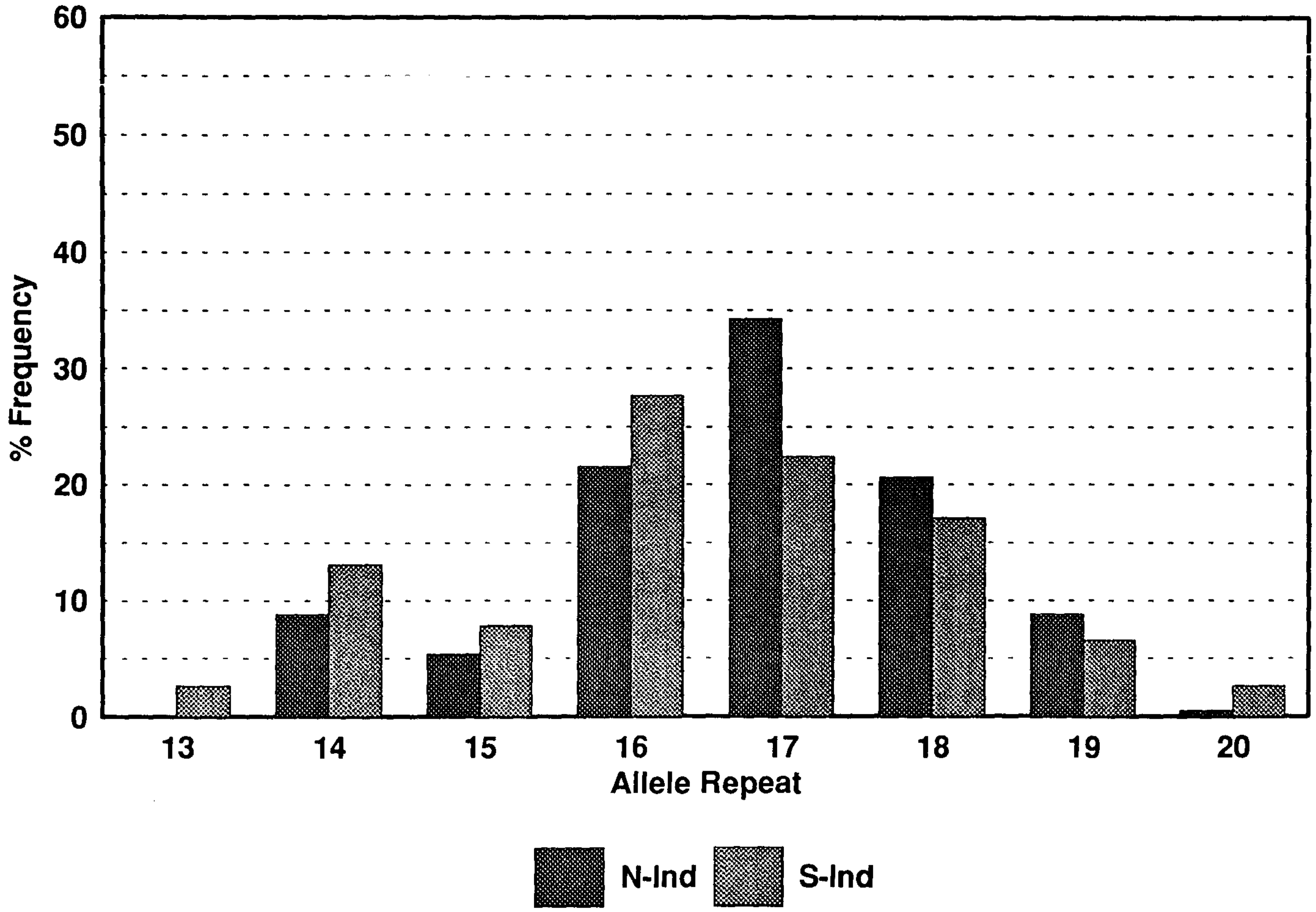


Table 8: Descriptive Statistics of the Allele Frequency Distributions of North and South India at Six Microsatellite Loci.

	Populations	
Locus	**North India**	**South India**
CSF1PO		
n	111	37
Mean	11.18 +/- 1.06	11.27 +/- 0.95
Mode	12	11
Min. allele	8	9
Max. allele	15	14
Median	11.5	11.5
# alleles	8	6
TPOX		
n	111	37
Mean	9.40 +/- 1.36	9.73 +/- 1.43
Mode	8	11
Min. allele	8	8
Max. allele	12	13
Median	10	10.5
# alleles	5	6
THO1		
n	111	37
Mean	7.79 +/- 1.46	7.92 +/- 1.43
Mode	6	9
Min. allele	6	6
Max. allele	10	10
Median	8	8
# alleles	5	5
F13A01		
n	95	30
Mean	6.01 +/- 2.59	6.07 +/- 2.95
Mode	5	5
Min. allele	4	4
Max. allele	16	16
Median	9.5	10
# alleles	8	8
FESFPS		
n	94	30
Mean	11.37 +/- 0.96	11.33 +/- 0.98
Mode	11	11
Min. allele	9	10
Max. allele	13	13
Median	11	11.5
# alleles	5	4
vWA		
n	103	37
Mean	16.82 +/- 1.34	16.41 +/- 1.55
Mode	17	16
Min. allele	14	13
Max. allele	20	20
Median	17	16.5
# alleles	7	8

was also different at four (CSF1PO, TPOX, FESFPS, vWA) out of the six loci between North and South India. Furthermore, at two loci (TPOX, vWA), the South Indian population which has a sample size 1/3 of that of the North Indian population contained alleles that were not present in the larger population. The alleles that were present in the South Indian population, was allele 13 at the TPOX locus and allele 13, at the vWA locus.

Relationship between India and Other World Populations

Homogeneity tests for pairwise contingency tables of allele frequency distributions were performed between the Indian, and the African, Caucasian and Hispanic populations at six microsatellite loci and between the Indian and the Mexican and Asian populations at four microsatellite loci. **Tables 9 - 14** show the results of the contingency analysis and **figures 14 - 19** show histograms comparing the allele frequency distributions of these populations at six microsatellite loci. Chi-square, G-test and Monte-Carlo simulations revealed significant differences ($P < 0.05$) in the allele frequency distributions between the Indian and African populations at all six microsatellite loci. Comparisons between the Indian and Caucasian populations revealed significant deviations from homogeneity at the TPOX, TH01, F13A01 and FESFPS loci, and the Hispanic population when compared to the Indian, had significant differences in allele frequency distributions at the at the TPOX, TH01 and F13A01 loci. When the Mexican and Asian populations were compared to the Indian population, at four microsatellite loci, the TH01 and F13A01 loci revealed allele frequency distributions that were significantly different.

A summary of the contingency analysis is shown in **table 15**. When six loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA) were considered in the contingency analysis, allele frequency distributions of three populations (African, Caucasian, Hispanic) were compared to that of the Indian population. It was shown that the allele frequency distributions of the African, Caucasian and Hispanic populations were significantly different ($P < 0.05$) from that of the Indian

Figure 14: Histograrn of allele frequency distribution.

Histograrn of allele frequency distributions of the Indian (Ind), African (Afr), Caucasian (Cau), Hispanic (His), Mexican (Mex) and Asian (Asi) populations at the CSF1PO locus.

Allele Frequency Distribution at the CSF1PO Locus

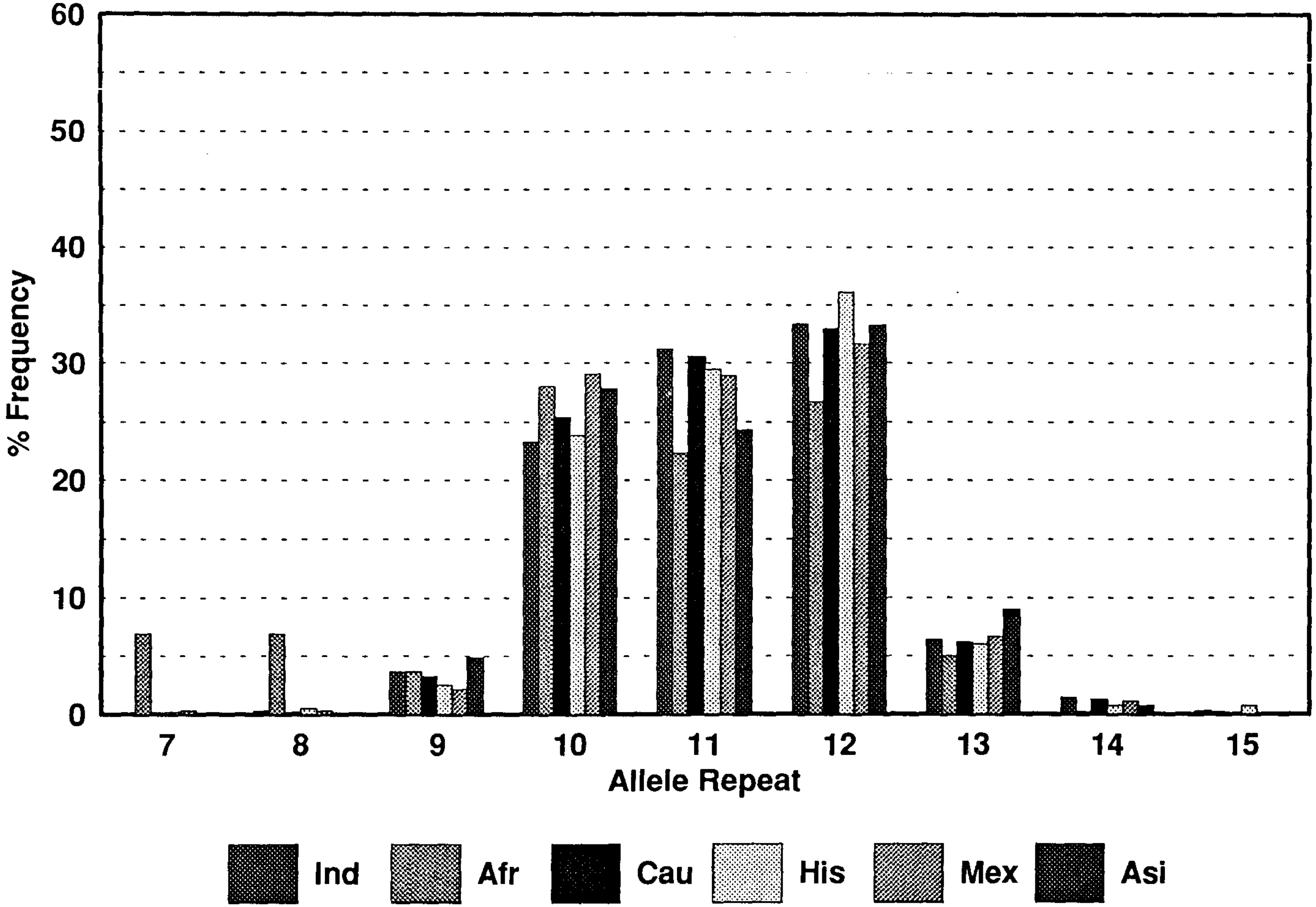


Table 9: Pairwise Contingency Table Analysis of Allele Frequency Distributions for India and Other World Populations at the CSF1PO Locus.

	Statistic					
Population	**X^2 - Test**		**G - Test**		**M.C. - Test**	
African	X^2	= 37.3	G	= 41.4	**P*	= 0
	df	= 4	*df*	= 4		
	**P*	< 0.05	**P*	< 0.05		
Caucasian	X^2	= 0.73	G	= 0.60	*P*	= 0.993
	df	= 5	*df*	= 5		
	P	> 0.05	*P*	> 0.05		
Hispanic	X^2	= 1.02	G	= 0.80	*P*	= 0.946
	df	= 5	*df*	= 5		
	P	> 0.05	*P*	> 0.05		
Mexican	X^2	= 3.55	G	= 3.80	*P*	= 0.664
	df	= 4	*df*	= 4		
	P	> 0.05	*P*	> 0.05		
Asian	X^2	= 2.76	G	= 2.80	*P*	= 0.709
	df	= 4	*df*	= 4		
	P	> 0.05	*P*	> 0.05		

* Significant P < 0.05.

Figure 15: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Indian (Ind), African (Afr), Caucasian (Cau), and Hispanic (His) populations at the TPOX locus.

Allele Frequency Distribution at the TPOX Locus

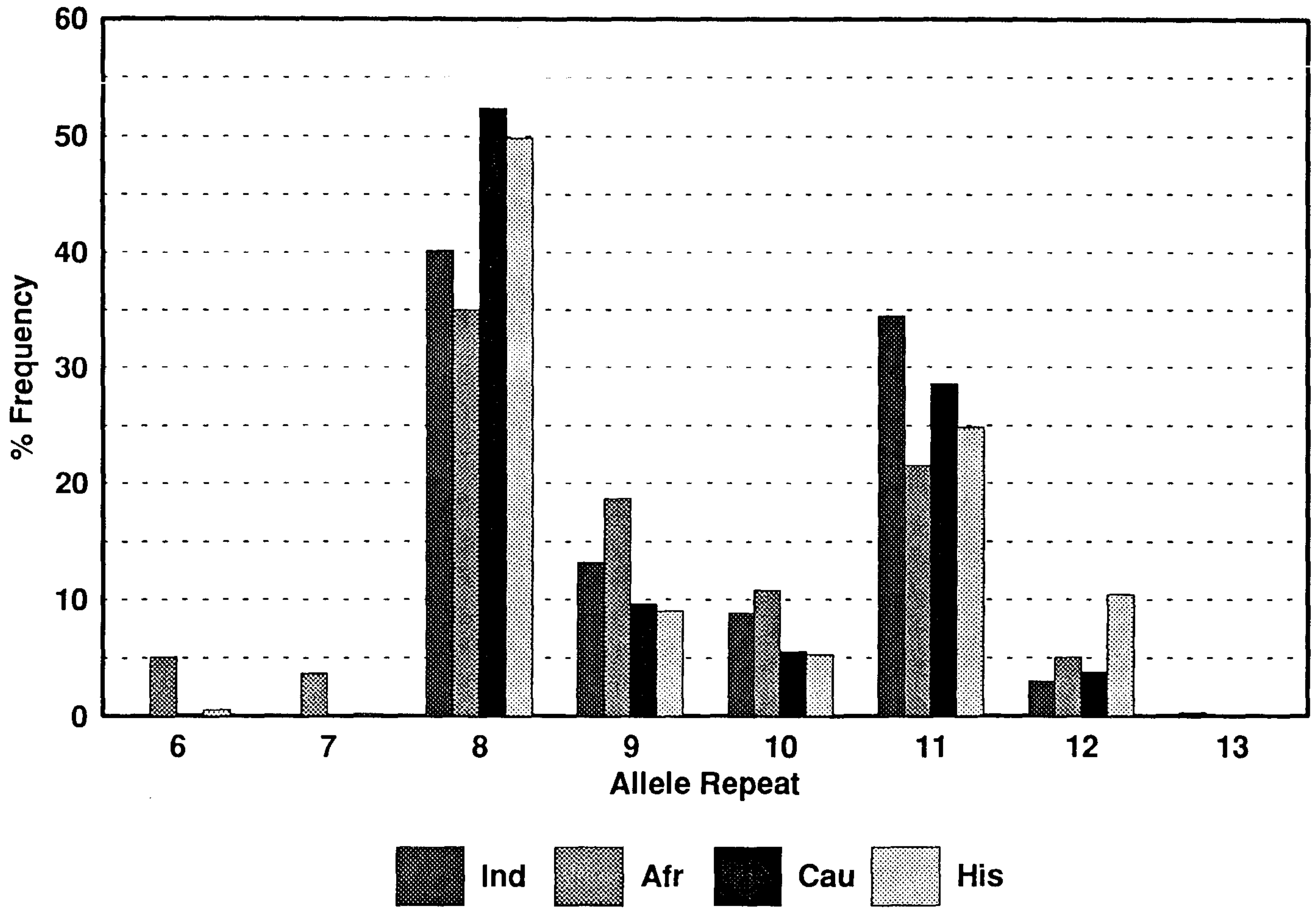


Table 10: Pairwise Contingency Table Analysis of Allele Frequency Distributions for India and Other World Populations at the TPOX Locus.

	Statistic		
Population	**X^2 - Test**	**G - Test**	**M.C. - Test**
African	X^2 = 16.1	G = 15.8	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	
Caucasian	X^2 = 12.5	G = 12.6	*P = 0.017
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	
Hispanic	X^2 = 27.2	G = 28.2	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	

* Significant $P < 0.05$.

Figure 16: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Indian (Ind), African (Afr), Caucasian (Cau), Hispanic (His), Mexican (Mex) and Asian (Asi) populations at the TH01locus.

Allele Frequency Distribution at the Th01 Locus

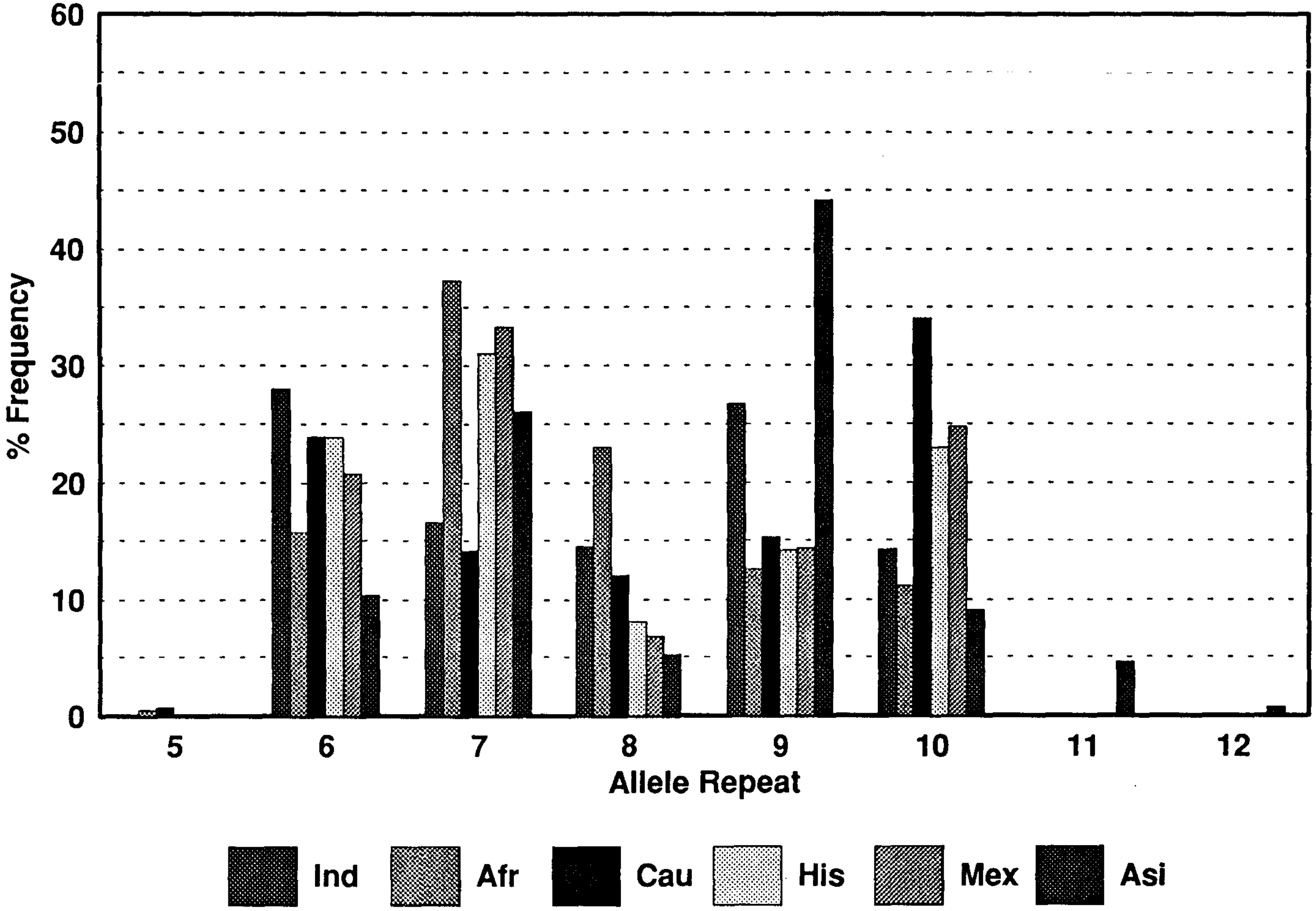


Table 11: Pairwise Contingency Table Analysis of Allele Frequency Distributions for India and Other World Populations at the TH01 Locus.

	Statistic		
Population	**X^2 - Test**	**G - Test**	**M.C. - Test**
African	X^2 = 63.6	G = 65.2	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	
Caucasian	X^2 = 39.8	G = 42.0	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	
Hispanic	X^2 = 43.9	G = 44.6	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	
Mexican	X^2 = 53.8	G = 55.6	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	
Asian	X^2 = 36.1	G = 38.8	*P = 0
	df = 4	df = 4	
	*P < 0.05	*P < 0.05	

* Significant $P < 0.05$.

Figure 17: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Indian (Ind), African (Afr), Caucasian (Cau), Hispanic (His), Mexican (Mex) and Asian (Asi) populations at the F13A01locus.

Allele Frequency Distribution at the F13A01 Locus

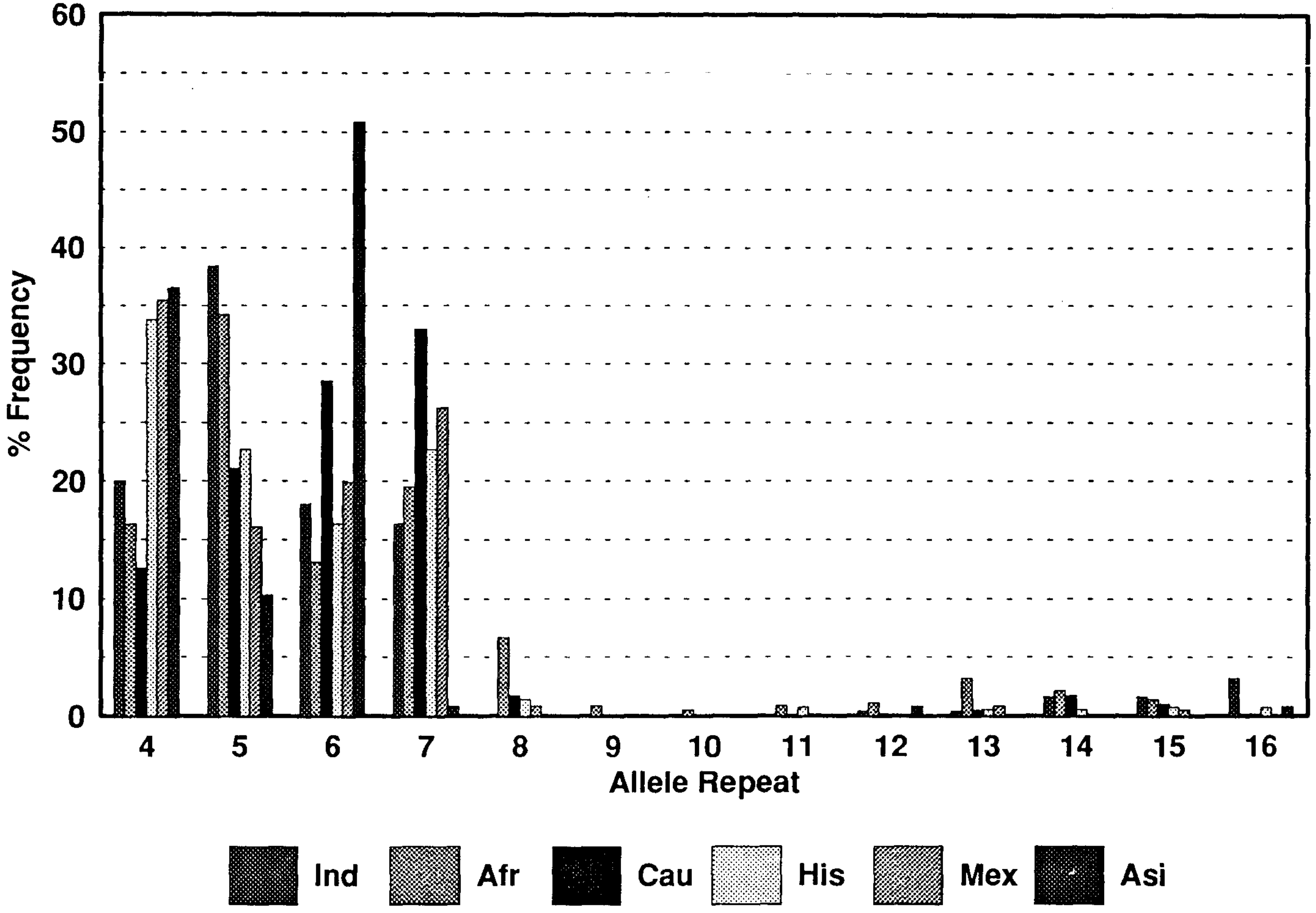


Table 12: Pairwise Contingency Table Analysis of Allele Frequency Distributions for India and Other World Populations at the F13A01 Locus.

	Statistic		
Population	**X^2 - Test**	**G - Test**	**M.C. - Test**
African	X^2 = 36.4	G = 42.6	*P = 0
	df = 5	*df* = 5	
	*P < 0.05	*P < 0.05	
Caucasian	X^2 = 47.3	G = 48.6	*P = 0
	df = 4	*df* = 4	
	*P < 0.05	*P < 0.05	
Hispanic	X^2 = 35.2	G = 35.4	*P = 0
	df = 5	*df* = 5	
	*P < 0.05	*P < 0.05	
Mexican	X^2 = 57.1	G = 57.4	*P = 0
	df = 4	*df* = 4	
	*P < 0.05	*P < 0.05	
Asian	X^2 = 34.2	G = 37.6	*P = 0
	df = 2	*df* = 2	
	*P < 0.05	*P < 0.05	

* Significant $P < 0.05$.

Figure 18: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Indian (Ind), African (Afr), Caucasian (Cau), Hispanic (His), Mexican (Mex) and Asian (Asi) populations at the FESFPS locus.

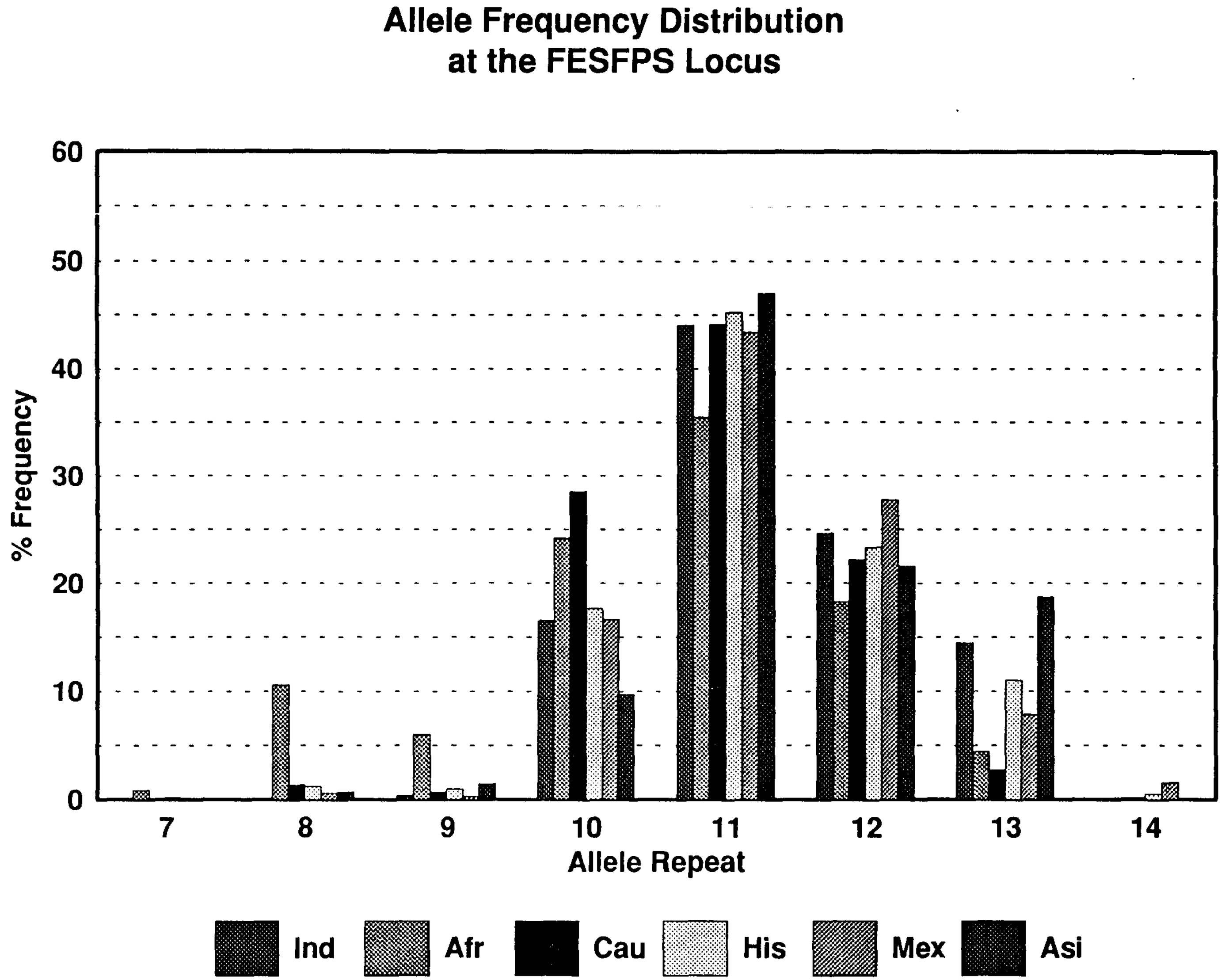

Allele Frequency Distribution
at the FESFPS Locus
60
50
40
30
20
10
0
% Frequency
7
8
9
10
11
12
13
14
Allele Repeat
Ind
Afr
Cau
His
Mex
Asi

Table 13: Pairwise Contingency Table Analysis of Allele Frequency Distributions for India and Other World Populations at the FESFPS Locus.

	Statistic		
Population	**X^2 - Test**	**G - Test**	**M.C. - Test**
African	X^2 = 54.6	G = 57.0	*P = 0
	df = 3	*df* = 3	
	*P < 0.05	*P < 0.05	
Caucasian	X^2 = 40.1	G = 40.0	*P = 0
	df = 3	*df* = 3	
	*P < 0.05	*P < 0.05	
Hispanic	X^2 = 2.12	G = 2.40	P = 0.402
	df = 3	*df* = 3	
	P > 0.05	P > 0.05	
Mexican	X^2 = 3.57	G = 4.20	P = 0.056
	df = 3	*df* = 3	
	P > 0.05	P > 0.05	
Asian	X^2 = 2.87	G = 3.20	P = 0.171
	df = 3	*df* = 3	
	P > 0.05	P > 0.05	

* Significant $P < 0.05$.

Figure 19: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Indian (Ind), African (Afr), Caucasian (Cau), and Hispanic (His) populations at the vWA locus.

Allele Frequency Distribution
at the vWA Locus

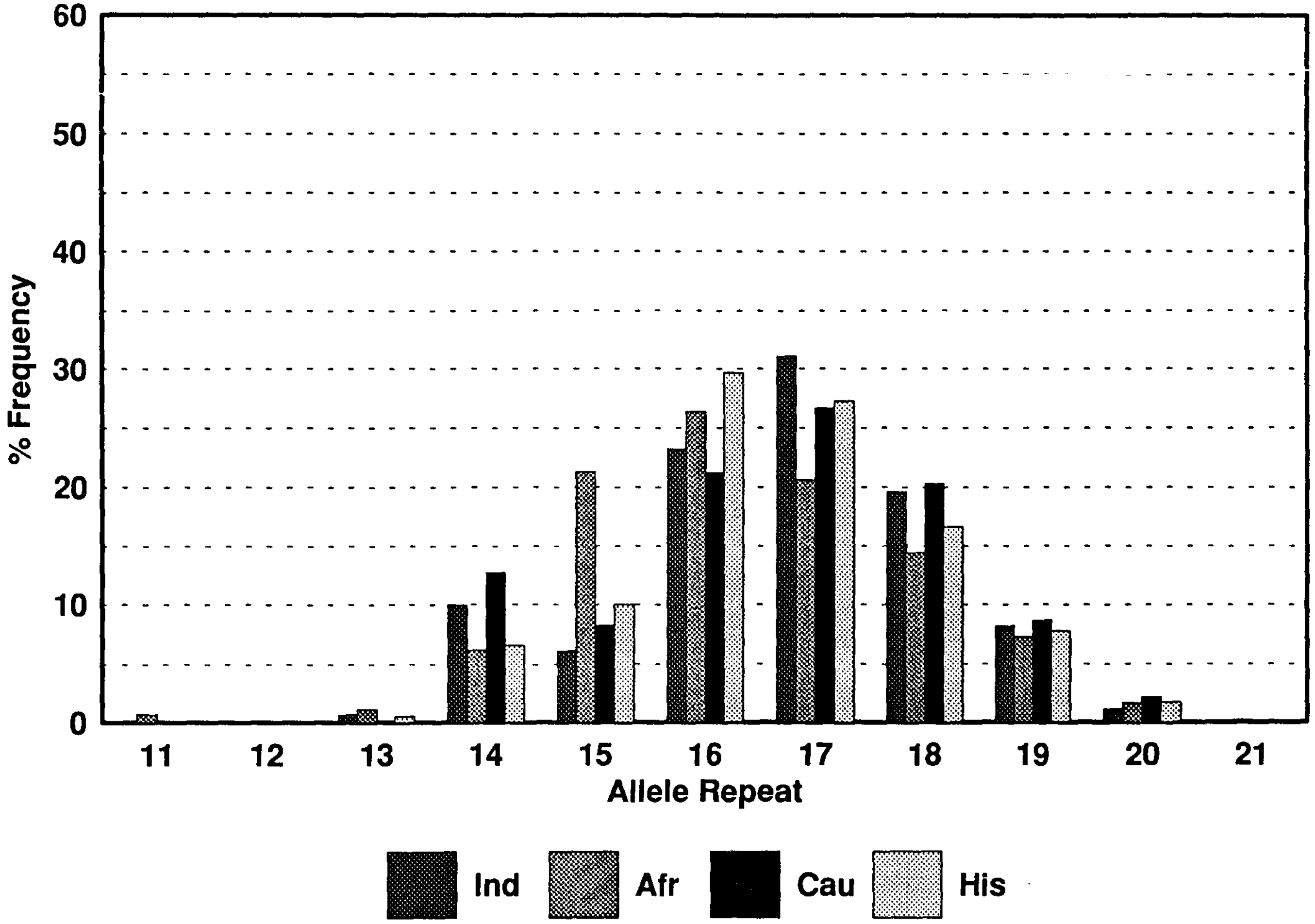


Table 14: Pairwise Contingency Table Analysis of Allele Frequency Distributions for India and Other World Populations at the vWA Locus.

	Statistic		
Population	**X^2 - Test**	**G - Test**	**M.C. - Test**
African	X^2 = 38.2	G = 41.6	**P* = 0
	df = 5	*df* = 5	
	**P* < 0.05	**P* < 0.05	
Caucasian	X^2 = 3.57	G = 3.80	*P* = 0.377
	df = 5	*df* = 5	
	P > 0.05	*P* > 0.05	
Hispanic	X^2 = 9.87	G = 9.80	*P* = 0.160
	df = 5	*df* = 5	
	P > 0.05	*P* > 0.05	

* Significant P < 0.05.

Table 15: Summary of Pairwise Contingency Analysis between India and Other World Populations at Six Microsatellite Loci.

	Locus							
Population	CSF1PO	TPOX	TH01	F13A01	FESFPS	vWA	% s - 6 loci	% s - 4 loci
African	s	s	s	s	s	s	100	100
Caucasian	n/s	s	s	s	s	n/s	67	75
Hispanic	n/s	s	s	s	n/s	n/s	50	50
Mexican	n/s	/	s	s	n/s	/	/	50
Asian	n/s	/	s	s	n/s	/	/	50
% s - 5 pops	20	/	100	100	40	/		
% s - 3 pops	33	100	100	100	67	33		

population at 100%, 67% and 50% of the loci respectively. When four loci (CSF1PO, TH01, F13A01, FESFPS) were considered, five populations (African, Caucasian, Hispanic, Mexican, Asian) were compared to that of the Indian population in the contingency analysis. It was shown that the allele frequency distributions of the African, Caucasian, Hispanic, Mexican and Asian populations were significantly different from the Indian population at 100%, 75%, 50%, 50% and 50% of the loci respectively.

When examining each locus individually, the percentage of the total number of populations that had allele frequency distributions significantly different ($P < 0.05$) from that of the Indian population was determined. Using six loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA), three populations (African, Caucasian, Hispanic) were compared to the Indian population. At the CSF1PO locus, 33% of the populations had allele frequency distributions significantly different from that of the Indian population. The TPOX, TH01 and F13A01 loci had 100% of the populations with significant deviations from homogeneity of allele frequency distributions and the FESFPS and vWA loci had 67% and 33% of the populations yielding significant results.

When four loci (CSF1PO, TH01, F13A01, FESFPS) were considered, five populations (African, Caucasian, Hispanic, Mexican and Asian) were compared to the allele frequency distributions of the Indian population. The CSF1PO locus had 20% of the populations having significantly different allele frequency distributions from that of the Indian population, followed by 100% for the TH01 and F13A01 loci and 40% for the FESFPS locus.

Overall, the CSF1PO locus had the least number of populations with allele frequency distributions significantly different from that of the Indian population, and the allele frequency distributions of the African population was significantly different from the Indian population at the greatest number of loci.

Descriptive statistics were computed for all six populations at all six loci and are shown in **table 16**. The mean allele size of a particular locus over all populations was very close. The

Table 16: Descriptive Statistics of the Allele Frequency Distributions of Six Populations at Six Microsatellite Loci.

	Populations					
Locus	**Indian**	**African**	**Caucasian**	**Hispanic**	**Mexican**	**Asian**
CSF1PO						
n	148	202	209	216	187	72
Mean	11.20 +/- 1.06	10.50 +/- 1.57	11.16 +/- 1.06	11.22 +/- 1.05	11.13 +/- 1.03	11.16 +/- 1.09
Mode	12	10	12	12	12	12
Min. allele	8	7	8	7	7	9
Max. allele	15	15	14	15	14	14
Median	11.50	11	11	11	10.5	11.5
# alleles	8	9	7	9	8	6
TPOX						
n	148	204	209	216		
Mean	9.48 +/- 1.40	9.12 +/- 1.54	9.21 +/- 1.42	9.34 +/- 1.58		
Mode	8	8	8	8		
Min. allele	8	6	6	6		
Max. allele	13	12	12	12		
Median	10.5	9	9.5	9		
# alleles	6	7	6	7		
THO1						
n	148	204	209	216	192	77
Mean	7.82 +/- 1.47	7.64 +/- 1.25	8.19 +/- 1.63	7.81 +/- 1.53	7.89 +/- 1.50	8.32 +/- 1.36
Mode	6	7	10	7	7	9
Min. allele	6	5	5	6	6	6
Max. allele	10	10	10	10	10	12
Median	8	7.5	7.5	8	8	9
# alleles	5	6	6	5	5	7
F13A01						
n	125	218	209	222	183	63
Mean	6.02 +/- 2.69	6.36 +/- 2.52	6.17 +/- 1.81	5.58 +/- 1.90	5.51 +/- 1.59	5.30 +/- 1.48
Mode	5	5	7	4	4	6
Min. allele	4	4	4	4	4	4
Max. allele	16	16	15	16	15	16
Median	12	10	8	9.5	7	6.5
# alleles	9	13	9	10	7	6
FESFPS						
n	124	217	212	210	159	67
Mean	11.36 +/- 0.97	10.60 +/- 0.96	10.94 +/- 0.93	11.23 +/- 0.97	11.08 +/- 1.81	11.36 +/- 1.39
Mode	11	11	11	11	11	11
Min. allele	9	7	8	7	8	8
Max. allele	13	14	14	14	14	13
Median	11	10.5	11	10.5	11	10.5
# alleles	5	8	7	8	7	6
vWA						
n	140	218	212	211		
Mean	16.71 +/- 1.43	16.39 +/- 1.48	16.68 +/- 1.54	16.66 +/- 1.36		
Mode	17	16	17	16		
Min. allele	13	11	14	13		
Max. allele	20	21	20	20		
Median	16.5	16.5	17	16.5		
# alleles	8	10	7	8		

features of the distributions that differed the most between populations, over all the loci, was the mode and the total number of alleles. At two loci (TPOX, FESFPS), the mode was the same for all of the populations. When looking at each locus individually, it was revealed at the CSF1PO locus, that the African population had a different mode (10) than the rest of the populations (12). The TPOX locus revealed that the African, Caucasian and Hispanic populations had a minimum allele size of 6, however, the Indian population's minimum allele size was 8. Furthermore, the maximum allele size of the three former populations was 12, and for the Indian population was 13. The mode at the TH01 locus was lowest in the Indian population compared to the others. The F13A01 locus displayed wide ranging median values (6.5 - 12). Although the minimum and maximum alleles were the same for most of the populations, the number of alleles found in between, were not consistent for all of the populations. As a result, the number of alleles found at this locus differs widely across all of the populations. The FESFPS locus displayed the same mode for all populations, however, the total number of alleles found at this locus was different for many of the populations. As well, the mean allele size for the African and Caucasian populations was slightly lower than the rest of the populations. The allele frequency distributions at the vWA locus revealed that the African population contained the most number of alleles, some of which were not found in other populations.

Relationship between Four Regions within India

Pairwise contingency analysis of allele frequency distributions between the four regions (Northcentral, Northeast, Southeast, Southwest) within India did not reveal any significant results. Only one pairwise comparison, between Northeast and Southwest India revealed a significant difference ($P < 0.05$) in allele frequency distribution at the CSF1PO locus using the Monte-Carlo simulation ($P = 0.005$). The results of the χ^2 and G-test were not significant. The results of the contingency analysis are shown in **tables 17 - 22** and histograms of the allele

Figure 20: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Northcentral (NC),Northeastern (NE), Southeastern (SE) and Southwestern (SW) populations within India at the CSF1PO locus.

Allele Frequency Distribution at the CSF1PO Locus Within India

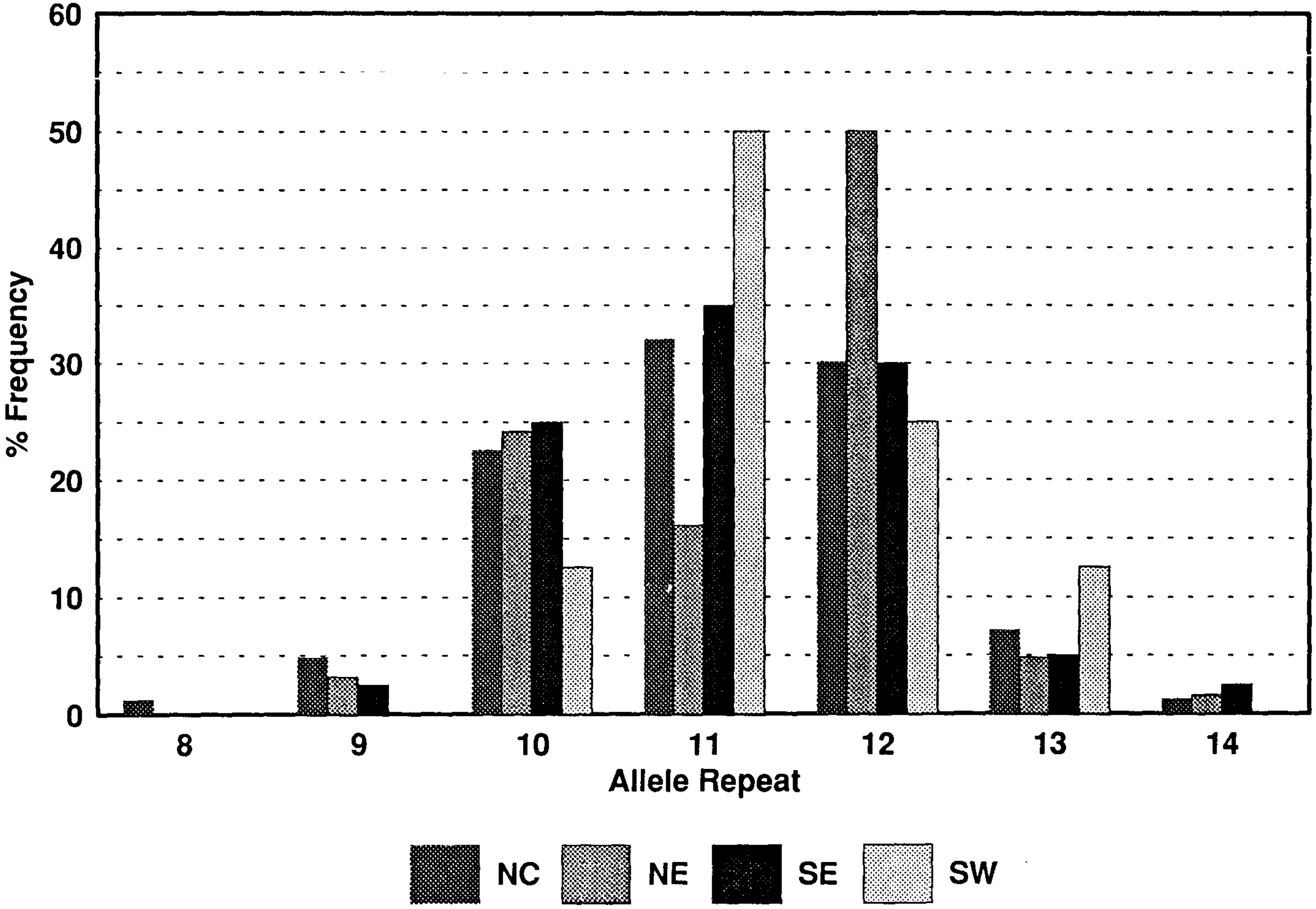


Table 17: Pairwise Contingency Table Analysis of Allele Frequency Distributions between Four Regions Within India at the CSF1PO Locus.

	Statistic		
Populations	**X^2 - Test**	**G - Test**	**M.C. - Test**
NC vs. NE	X^2 = 5.88 *df* = 2 *P* > 0.05	G = 5.94 *df* = 2 *P* > 0.05	*P* = 0.188
NC vs. SE	X^2 = 0.10 *df* = 2 *P* > 0.05	G = 0 *df* = 2 *P* > 0.05	*P* = 0.993
NC vs. SW	X^2 = 0.03 *df* = 1 *P* > 0.05	G = 0 *df* = 1 *P* > 0.05	*P* = 0.381
NE vs. SE	X^2 = 5.46 *df* = 2 *P* > 0.05	G = 5.00 *df* = 2 *P* > 0.05	*P* = 0.319
NE vs. SW	X^2 = 3.03 *df* = 1 *P* > 0.05	G = 2.80 *df* = 1 *P* > 0.05	**P* = 0.005
SE vs. SW	X^2 = 0 *df* = 1 *P* > 0.05	G = 0 *df* = 1 *P* > 0.05	*P* = 0.321

* Significant $P < 0.05$.

Figure 21: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Northcentral (NC), Northeastern (NE), Southeastern (SE) and Southwestern (SW) populations within India at the TPOX locus.

Allele Frequency Distribution at the TPOX Locus Within India

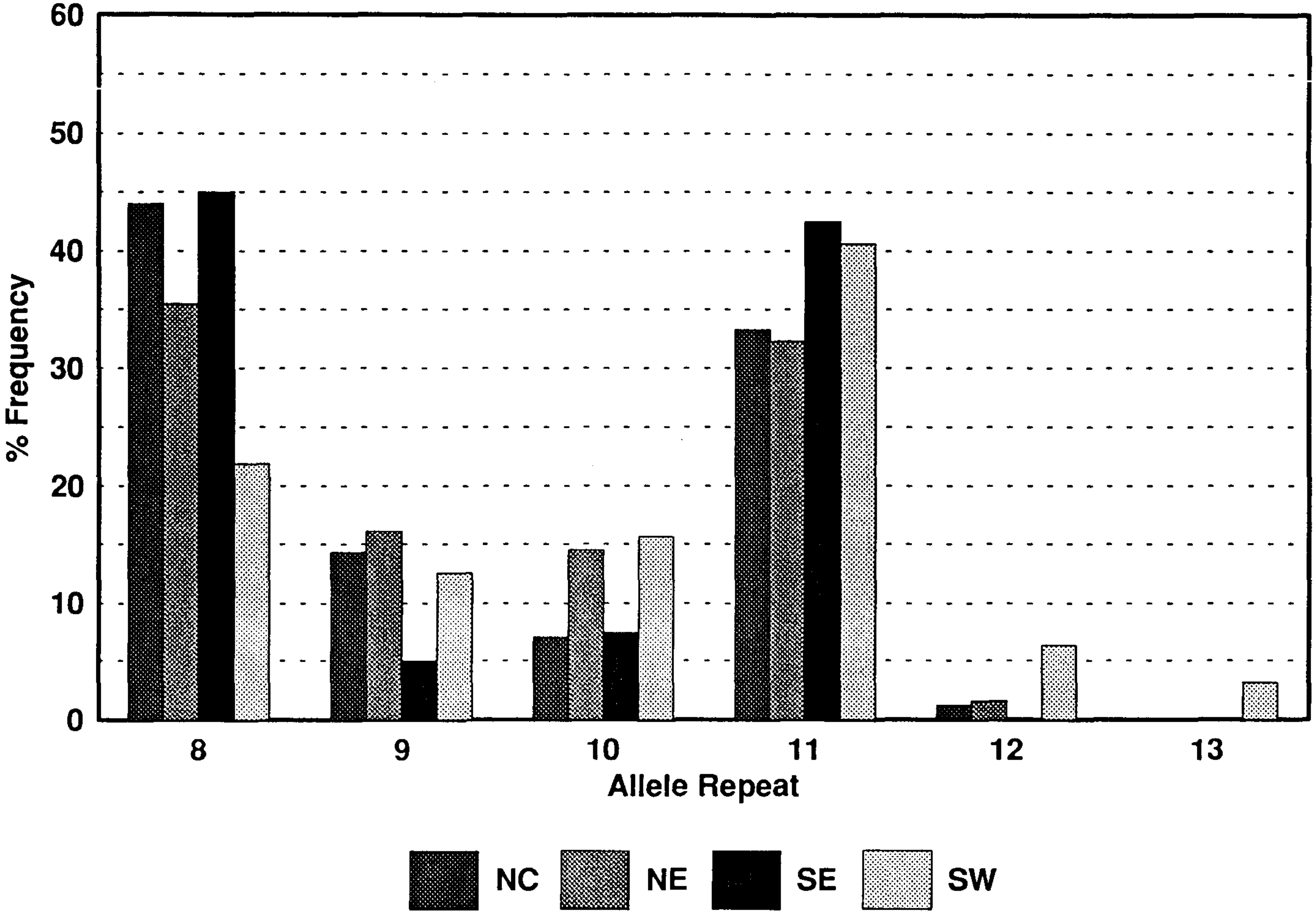


Table 18: Pairwise Contingency Table Analysis of Allele Frequency Distributions between Four Regions Within India at the TPOX Locus.

	Statistic		
Populations	**X^2 - Test**	**G - Test**	**M.C. - Test**
NC vs. NE	X^2 = 2.62	G = 2.80	P = 0.663
	df = 3	df = 3	
	P > 0.05	P > 0.05	
NC vs. SE	X^2 = 1.63	G = 1.60	P = 0.524
	df = 2	df = 2	
	P > 0.05	P > 0.05	
NC vs. SW	X^2 = 4.88	G = 5.20	P = 0.061
	df = 2	df = 2	
	P > 0.05	P > 0.05	
NE vs. SE	X^2 = 4.45	G = 4.20	P = 0.228
	df = 2	df = 2	
	P > 0.05	P > 0.05	
NE vs. SW	X^2 = 2.71	G = 2.60	P = 0.388
	df = 2	df = 2	
	P > 0.05	P > 0.05	
SE vs. SW	X^2 = 5.19	G = 5.20	P = 0.079
	df = 2	df = 2	
	P > 0.05	P > 0.05	

Figure 22: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Northcentral (NC), Northeastern (NE), Southeastern (SE) and Southwestern (SW) populations within India at the TH01 locus.

Allele Frequency Distribution at the TH01 Locus Within India

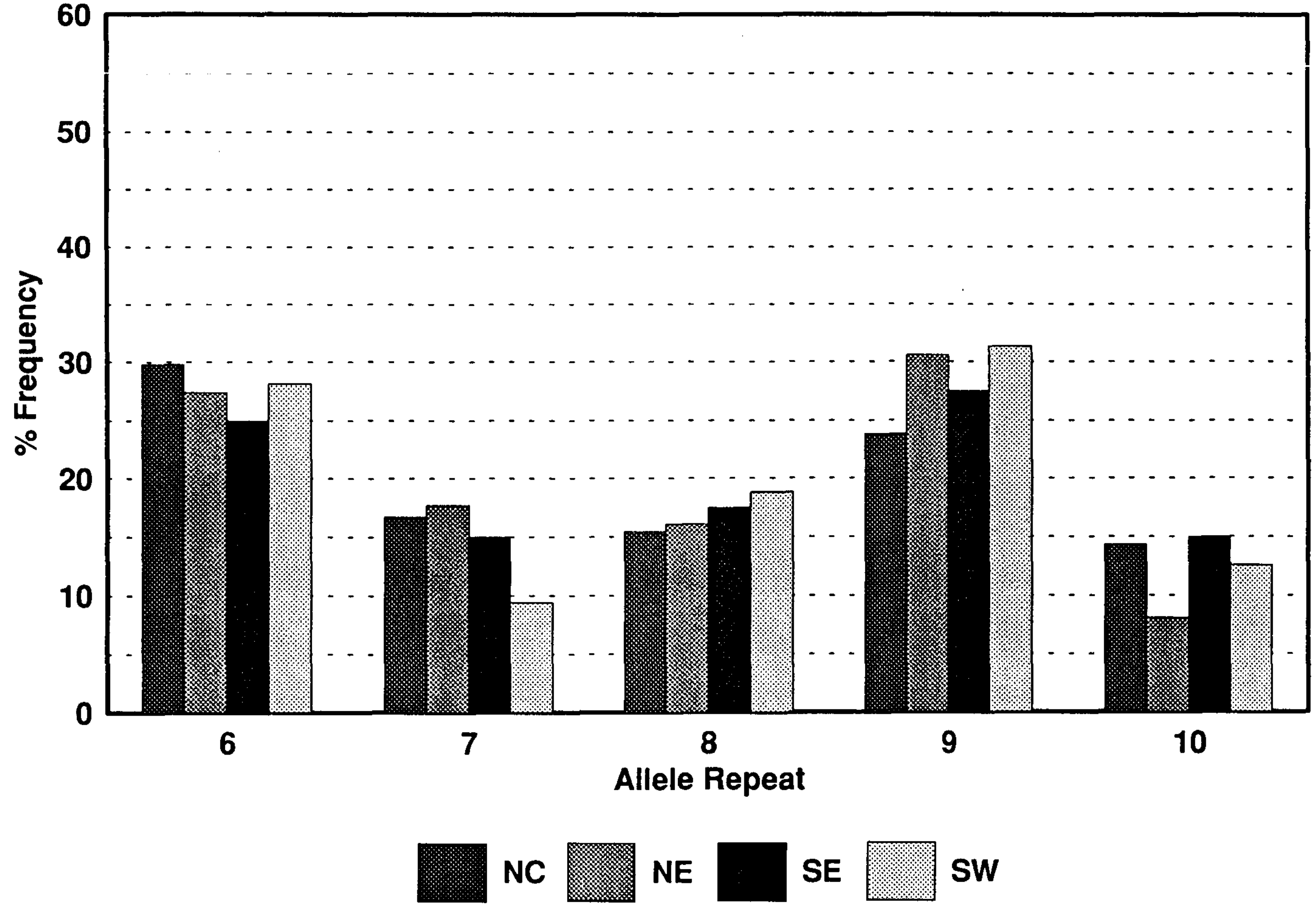


Table 19: Pairwise Contingency Table Analysis of Allele Frequency Distributions between Four Regions Within India at the TH01 Locus.

	Statistic		
Populations	**X^2 - Test**	**G - Test**	**M.C. - Test**
NC vs. NE	X^2 = 1.91 *df* = 4 *P* > 0.05	G = 1.40 *df* = 4 *P* > 0.05	*P* = 0.763
NC vs. SE	X^2 = 0.49 *df* 4 *P* > 0.05	G = 0.40 *df* = 4 *P* > 0.05	*P* = 0.980
NC vs. SW	X^2 = 0.33 *df* = 2 *P* > 0.05	G = 0.40 *df* = 2 *P* > 0.05	*P* = 0.821
NE vs. SE	X^2 = 1.36 *df* = 4 *P* > 0.05	G = 1.00 *df* = 4 *P* > 0.05	*P* = 0.870
NE vs. SW	X^2 = 0.36 *df* = 2 *P* > 0.05	G = 0.60 *df* = 2 *P* > 0.05	*P* = 0.832
SE vs. SW	X^2 = 0.18 *df* = 2 *P* > 0.05	G = 0 *df* = 2 *P* > 0.05	*P* = 0.959

Figure 23: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Northcentral (NC), Northeastern (NE), Southeastern (SE) and Southwestern (SW) populations within India at the F13A01 locus.

Allele Frequency Distribution at the F13A01 Locus Within India

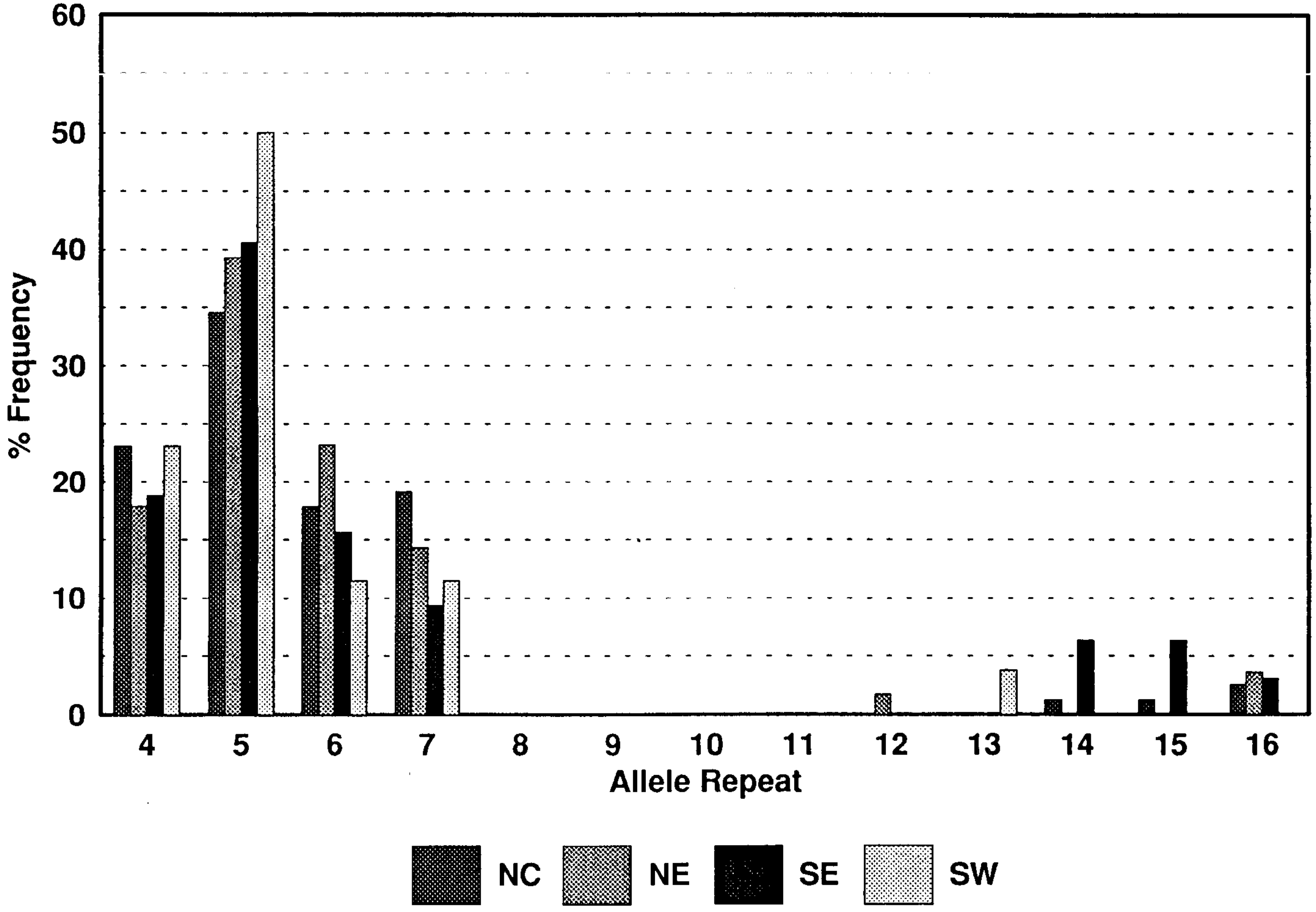


Table 20: Pairwise Contingency Table Analysis of Allele Frequency Distributions between Four Regions Within India at the F13A01 Locus.

	Statistic		
Populations	**X^2 - Test**	**G - Test**	**M.C. - Test**
NC vs. NE	X^2 = 1.39	G = 1.20	P = 0.839
	df = 3	df = 3	
	P > 0.05	P > 0.05	
NC vs. SE	X^2 = 0.50	G = 0.20	P = 0.412
	df = 3	df = 3	
	P > 0.05	P > 0.05	
NC vs. SW	X^2 = 2.40	G = 2.20	P = 0.466
	df = 2	df = 2	
	P > 0.05	P > 0.05	
NE vs. SE	X^2 = 0.86	G = 0.80	P = 0.271
	df = 3	df = 3	
	P > 0.05	P > 0.05	
NE vs. SW	X^2 = 1.92	G = 2.00	P = 0.497
	df = 2	df = 2	
	P > 0.05	P > 0.05	
SE vs. SW	X^2 = 1.19	G = 1.20	P = 0.616
	df = 2	df = 2	
	P > 0.05	P > 0.05	

Figure 24: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Northcentral (NC), Northeastern (NE), Southeastern (SE) and Southwestern (SW) populations within Indiia at the FESFPS locus.

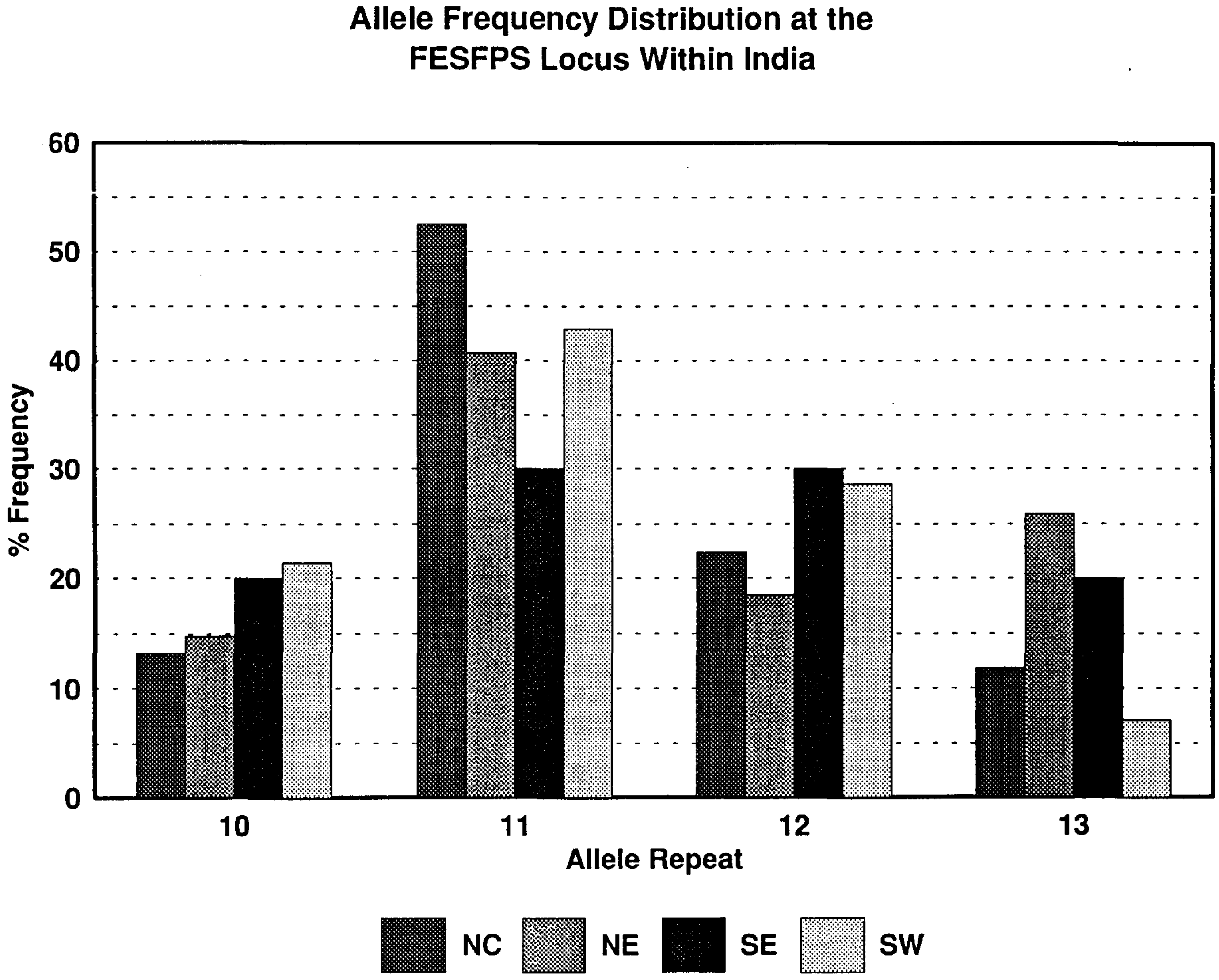

Allele Frequency Distribution at the
FESFPS Locus Within India
60
50
40
30
20
10
0
% Frequency
10
11
12
13
Allele Repeat
NC
NE
SE
SW

Table 21: Pairwise Contingency Table Analysis of Allele Frequency Distributions between Four Regions Within India at the FESFPS Locus.

	Statistic		
Populations	X^2 - Test	G - Test	M.C. - Test
NC vs. NE	$X^2 = 4.76$ $df = 3$ $P > 0.05$	$G = 4.80$ $df = 3$ $P > 0.05$	$P = 0.190$
NC vs. SE	$X^2 = 4.57$ $df = 3$ $P > 0.05$	$G = 5.20$ $df = 3$ $P > 0.05$	$P = 0.228$
NC vs. SW	$X^2 = 1.31$ $df = 2$ $P > 0.05$	$G = 1.20$ $df = 2$ $P > 0.05$	$P = 0.604$
NE vs. SE	$X^2 = 2.32$ $df = 3$ $P > 0.05$	$G = 2.60$ $df = 3$ $P > 0.05$	$P = 0.538$
NE vs. SW	$X^2 = 0.83$ $df = 2$ $P > 0.05$	$G = 1.00$ $df = 2$ $P > 0.05$	$P = 0.221$
SE vs. SW	$X^2 = 1.36$ $df = 2$ $P > 0.05$	$G = 1.20$ $df = 2$ $P > 0.05$	$P = 0.519$

Figure 25: Histogram of allele frequency distribution.

Histogram of allele frequency distributions of the Northcentral (NC), Northeastern (NE), Southeastern (SE) and Southwestern (SW) populations within India at the vWA locus.

Allele Frequency Distribution at the vWA Locus Within India

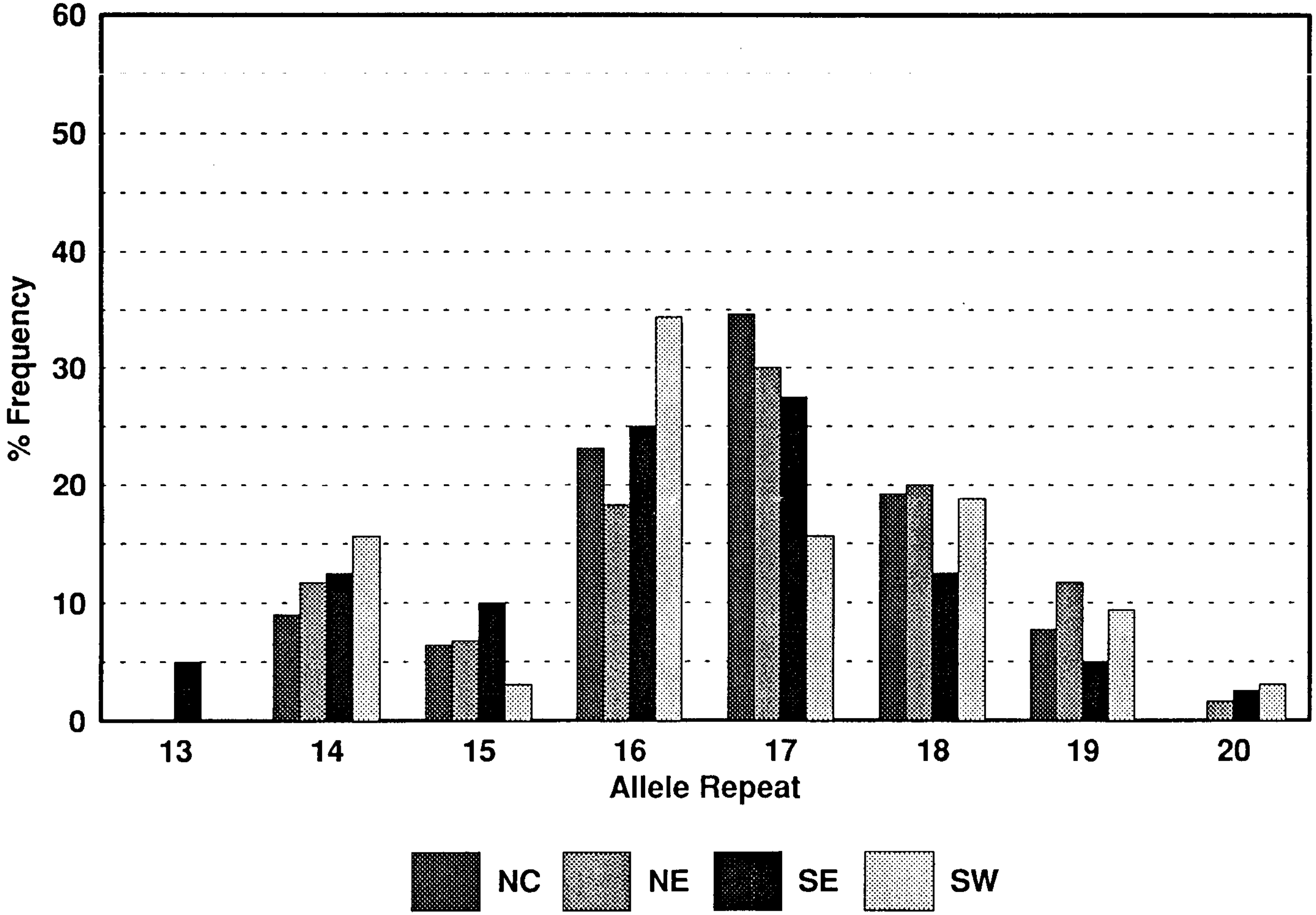


Table 22: Pairwise Contingency Table Analysis of Allele Frequency Distributions between Four Regions Within India at the vWA Locus.

		Statistic	
Populations	**X^2 - Test**	**G - Test**	**M.C. - Test**
NC vs. NE	$X^2 = 1.84$ $df = 4$ $P > 0.05$	G $= 1.80$ $df = 4$ $P > 0.05$	$P = 0.914$
NC vs. SE	$X^2 = 2.96$ $df = 3$ $P > 0.05$	G $= 2.60$ $df = 3$ $P > 0.05$	$P = 0.359$
NC vs. SW	$X^2 = 4.22$ $df = 3$ $P > 0.05$	G $= 4.40$ $df = 3$ $P > 0.05$	$P = 0.221$
NE vs. SE	$X^2 = 3.00$ $df = 3$ $P > 0.05$	G $= 2.80$ $df = 3$ $P > 0.05$	$P = 0.575$
NE vs. SW	$X^2 = 4.00$ $df = 3$ $P > 0.05$	G $= 4.00$ $df = 3$ $P > 0.05$	$P = 0.582$
SE vs. SW	$X^2 = 3.14$ $df = 3$ $P > 0.05$	G $= 3.20$ $df = 3$ $P > 0.05$	$P = 0.631$

frequency distributions of these four regions at six microsatellite loci are shown in **figures 20 - 25.**

Descriptive statistics shown on **table 23** of these four regions did not reveal any particular feature that was common or distinctive to any of the populations. Some loci displayed uniform results for all regions, however, other loci did not show any consistencies between the populations. The mean allele size was close among the four populations at each locus and most populations had the same mode at each locus. However, the Southeast population at the FESFPS locus displayed a bi-modal distribution, thought to be due to the small sample size. The one characteristic of the distributions that differed widely among the populations at most loci was the total number of alleles detected. However, this was not the case at the FESFPS and TH01 loci which had the same number of alleles for all regions.

Genetic Distance

Genetic distance was calculated using Goldsteins *et al.*, (1995) measure of $(\delta\mu)^2$. Genetic Distance was calculated using four and six microsatellite loci. When six microsatellite loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA) were utilised, the populations used in the analysis were Indian, African, Caucasian and Hispanic. **Table 24** shows genetic distance values (lower diagonal) and standard errors (upper diagonal). Genetic distance was recalculated using the same number of markers, however the Indian population was split into North and South regions. **Table 25** shows the genetic distances and standard errors of this comparison. Using four microsatellite markers (CSF1PO, TH01, F13A01, FESFPS), the genetic distance was calculated twice, as it was above, however, the Mexican and Asian populations were added to the analysis. **Tables 26 and 27** show genetic distances and standard errors obtained when four microsatellite loci were used.

Table 23: Descriptive Statistics of the Allele Frequency Distributions of Populations Within India at Six Microsatellite Loci.

	Populations			
Locus	NC India	NE India	SE India	SW India
CSF1PO				
n	42	31	20	16
Mean	11.13 +/- 1.11	11.34 +/- 1.03	11.18 +/- 0.97	11.38 +/- 0.79
Mode	11	12	11	11
Min. allele	8	9	9	10
Max. allele	14	14	14	13
Median	11	11.5	11.5	11.5
# alleles	7	6	6	4
TPOX				
n	42	31	20	16
Mean	9.33 +/- 1.38	9.48 +/- 1.33	9.48 +/- 1.38	10.06 +/- 1.41
Mode	8	8	8	11
Min. allele	8	8	8	8
Max. allele	12	12	11	13
Median	10	10	9.5	10.5
# alleles	5	5	4	6
THO1				
n	42	31	20	16
Mean	7.76 +/- 1.46	7.74 +/- 1.37	7.93 +/- 1.39	7.91 +/- 1.40
Mode	6	9	9	9
Min. allele	6	6	6	6
Max. allele	10	10	10	10
Median	8	8	8	8
# alleles	5	5	5	5
F13A01				
n	39	28	16	13
Mean	5.86 +/- 2.42	5.86 +/- 2.33	6.69 +/- 3.60	5.42 +/- 1.77
Mode	5	5	5	5
Min. allele	4	4	4	4
Max. allele	16	16	16	13
Median	7	6.5	7	6
# alleles	7	8	7	5
FESFPS				
n	38	27	15	14
Mean	11.33 +/- 0.83	11.56 +/- 0.98	11.50 +/- 1.02	11.21 +/- 0.91
Mode	11	11	11, 12	11
Min. allele	10	10	10	10
Max. allele	13	13	13	13
Median	11.5	11.5	11.5	11.5
# alleles	4	4	4	4
vWA				
n	39	30	20	16
Mean	16.72 +/- 1.28	16.82 +/- 1.47	16.28 +/- 1.58	16.59 +/- 1.64
Mode	17	17	17	16
Min. allele	14	14	13	14
Max. allele	19	20	20	20
Median	16.5	17	16.5	17
# alleles	6	7	8	7

Table 24: Genetic Distances (lower) and Standard Errors (upper) between Four Populations using Six Microsatellites.

	Indian	African	Caucasian	Hispanic
Indian		0.091	0.027	0.029
African	0.243*		0.058	0.095
Caucasian	0.068*	0.160*		0.050
Hispanic	0.039	0.276*	0.099	

* Significant at the 5% level.

Table 25: Genetic Distances (lower) and Standard Errors (upper) between Five Populations using Six Microsatellite Loci.

	N. Indian	S. Indian	African	Caucasian	Hispanic
N. Indian		0.028	0.092	0.030	0.027
S. Indian	0.052		0.098	0.037	0.036
African	0.245*	0.277*		0.062	0.095
Caucasian	0.071*	0.100*	0.160*		0.050
Hispanic	0.039	0.079*	0.276*	0.099	

* Significant at the 5% level.

Table 26: Genetic Distances (lower) and Standard Errors (upper) between Six Populations using Four Microsatellite Loci.

	Indian	African	Caucasian	Hispanic	Mexican	Asian
Indian		0.123	0.037	0.042	0.056	0.104
African	0.304*		0.068	0.109	0.125	0.157
Caucasian	0.084*	0.216*		0.063	0.081	0.156
Hispanic	0.054	0.382*	0.144*		0.001	0.037
Mexican	0.069	0.414*	0.162*	0.006*		0.027
Asian	0.181	0.665*	0.253	0.084*	0.052	

* Significant at the 5% level.

Table 27: Genetic Distances (lower) and Standard Errors (upper) between Seven Populations using Four Microsatellite Loci.

	N. Indian	S. Indian	African	Caucasian	Hispanic	Mexican	Asian
N. Indian		0.003	0.125	0.040	0.039	0.052	0.101
S. Indian	0.007*		0.119	0.029	0.049	0.065	0.118
African	0.300*	0.321*		0.068	0.109	0.125	0.157
Caucasian	0.092*	0.062*	0.216*		0.063	0.081	0.156
Hispanic	0.052	0.065	0.382*	0.144*		0.001	0.037
Mexican	0.066	0.082	0.414*	0.162*	0.006*		0.027
Asian	0.183	0.183	0.665*	0.253	0.084*	0.052	

* Significant at the 5% level.

The results of the genetic distance calculation revealed that the African population had genetic distance values that were always significant ($P < 0.05$) between all other populations when four and six microsatellite loci were used in the analysis. The Caucasian population had distance values that were always significant between all populations, except the Hispanic population when six microsatellite markers were used, and the Asian population when four microsatellite markers were used. The Indian population displayed distance values that were significant between the African and Caucasian population, but not the Hispanic, Mexican or Asian populations. When the Indian population was split into the Northern and Southern regions, the North Indian population had distance values that were significant between the African and Caucasian populations but never significant between the Hispanic, Mexican or Asian populations. The genetic relationship between North and South India when using six microsatellite markers revealed a distance value of ($(\delta\mu)^2 = 0.052$) which was not significant ($P > 0.05$). However, the use of four markers yielded a genetic distance of ($(\delta\mu)^2 = 0.007$) which was significant between these two regions. The South Indian population always had significant distance values between the African and Caucasian populations, but not with the Mexican or Asian populations. With six microsatellite markers, the genetic distance between the South Indian and Hispanic population was significant, but not with the use of four markers. Overall, the genetic distances separating the African population from the rest of the populations had the highest values compared to the others.

When considering only the genetic distances between India and the other populations, at four and six microsatellite loci, a trend was revealed between the genetic distance value and the percentage of loci that had a significant difference in allele frequency distributions between each population and the Indian population in the contingency analysis. **Tables 28 and 29** shows the results of this trend. For each population, the greater the number of loci that revealed a significant result, the higher the genetic distance value.

Table 28: Trend between Genetic Distance and the Percentage of Loci that have Significant Differences in the Allele Frequency Distributions from that of the Indian Population at Six Microsatellite Loci.

		Populations	
	African	**Caucasian**	**Hispanic**
Genetic Distance	0.243	0.068	0.039
% of Significant Loci	100	67	50

Table 29: Trend between Genetic Distance and the Percentage of Loci that have Significant Differences in the Allele Frequency Distributions from that of the Indian Population at Four Microsatellite Loci.

	Populations				
	African	**Caucasian**	**Hispanic**	**Mexican**	**Asian**
Genetic Distance	0.304	0.084	0.054	0.069	0.181
% of Significant Loci	100	75	50	50	50

Genetic distance was also computed between the four regions (Northcentral, Northeast, Southeast, Southwest) within India. **Table 30** shows the distance values (lower diagonal) and standard errors (upper diagonal). The genetic distance between the Northeast and Southwest region ($\delta\mu^2$ = 0.120) and between the Northeast and Northcentral region ($((\delta\mu)^2$ = 0.021) was significant ($P < 0.05$). The other distance values were not statistically significant.

Phylogenetic Analysis

The Neighbor-Joining method of tree construction was used to infer the phylogenetic relationship between the Indian, African, Caucasian and Hispanic populations using six microsatellite markers (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA) and between the Indian, African, Caucasian, Hispanic, Mexican and Asian populations using four microsatellite markers (CSF1PO, TH01, F13A01, FESFPS). In both cases, the Indian population was split into the Northern and Southern regions and the phylogenetic relationship was inferred once again. **Figures 26 - 29** show the results of this analysis. In general, the cluster containing the Caucasian, Indian (North and South), Hispanic, Mexican and Asian populations was statistically supported at the 95% confidence level. Bootstrap values were close to or greater that 95% except in the situation where six microsatellites were used and the Indian population was not split into two regions (**see Figure 26**). In this case the cluster containing the Caucasian, Hispanic and Indian populations was not statistically supported, revealing a bootstrap value of 54%, however, the cluster containing the Indian and Hispanic populations was statistically significant with a bootstrap value of 95%. Overall, the rest of the branches were not statistically supported at the 95% confidence level as indicated by the bootstrap values in all four trees. However, almost all had bootstrap values over 50%.

The results of the phylogenetic analysis when using four or six microsatellite markers revealed the same branching order or topology of the populations. When the Indian population

Table 30: Genetic Distances (lower) and Standard Errors (upper) between Four Populations Within India using Six Microsatellite Loci.

	Northeast	Northcentral	Southeast	Southwest
Northeast		0.008	0.153	0.048
Northcentral	0.021*		0.151	0.077
Southeast	0.229	0.215		0.309
Southwest	0.120*	0.138	0.442	

* Significant at the 5% level.

Figure 26: Phylogenetic tree inferred using the Neighbor-Joining algorithm (Saitou and Nei, 1987). Tree was generated by the software package PHYLIP and labels were plotted using the software package TREETOOL.

Phylogenetic tree of four populations - Indian (Ind), African (Afr), Caucasian (Cau) and Hispanic (His) using six microsatellite markers. Bootstrap values generated using the CONSENSE program were placed at the nodes.

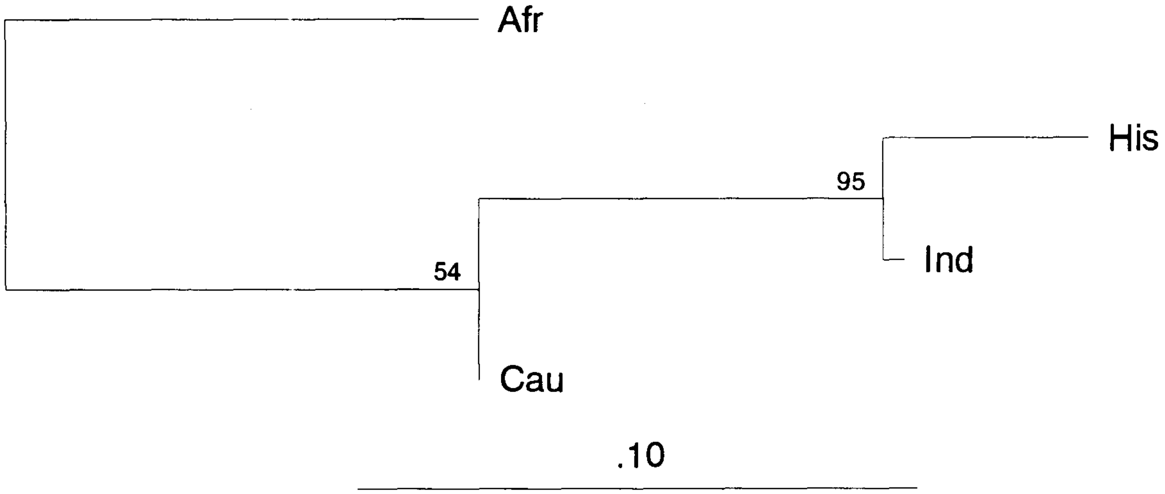

Afr
His
95
Ind
54
Cau
.10

Figure 27: Phylogenetic tree inferred using the Neighbor-Joining algorithm (Saitou and Nei, 1987). Tree was generated by the software package PHYLIP and labels were plotted using the software package TREETOOL.

Phylogenetic tree of five populations - North Indian (N. Ind), South Indian (S. Ind), African (Afr), Caucasian (Cau) and Hispanic (His) using six microsatellite markers. Bootstrap values generated using the CONSENSE program were placed at the nodes.

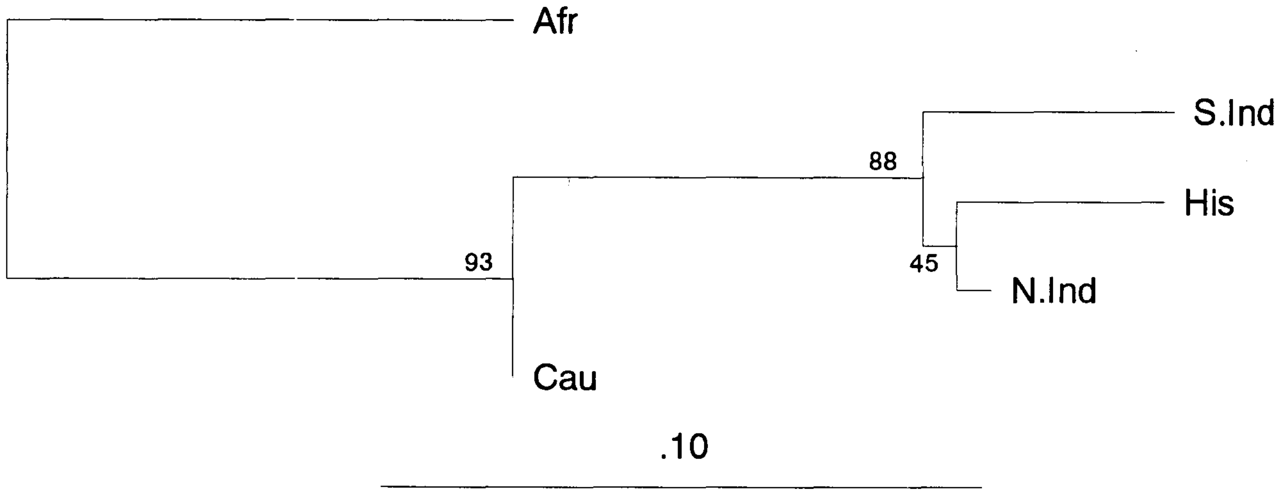

Afr
S.Ind
88
His
93
45
N.Ind
Cau
.10

Figure 28: Phylogenetic tree inferred using the Neighbor-Joining algorithm (Saitou and Nei, 1987). Tree was generated by the software package PHYLIP and labels were plotted using the software package TREETOOL.

Phylogenetic tree of six populations - Indian (Ind), African (Afr), Caucasian (Cau), Hispanic (His), Mexican (Mex) and Asian (Asi) using four microsatellite markers. Bootstrap values generated using the CONSENSE program were placed at the nodes.

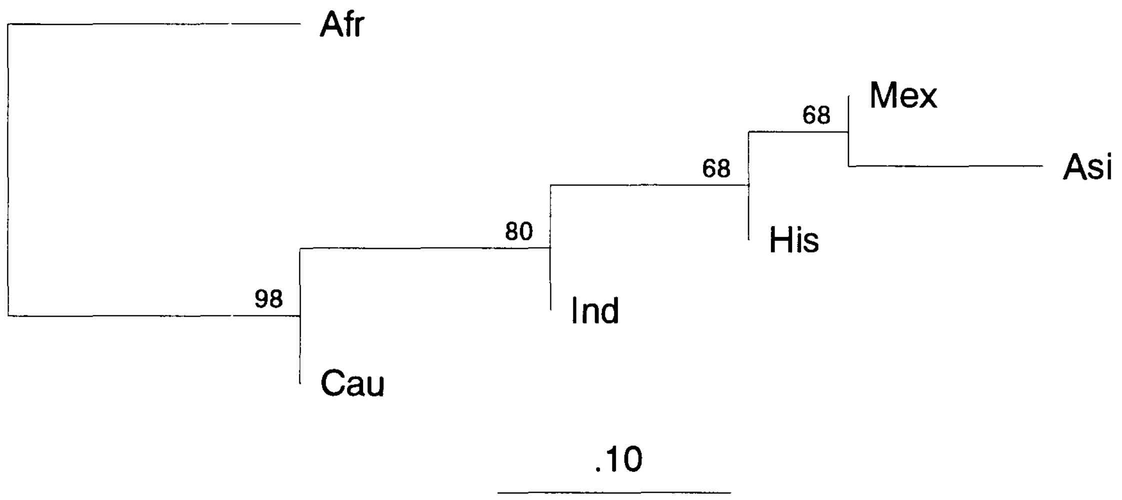
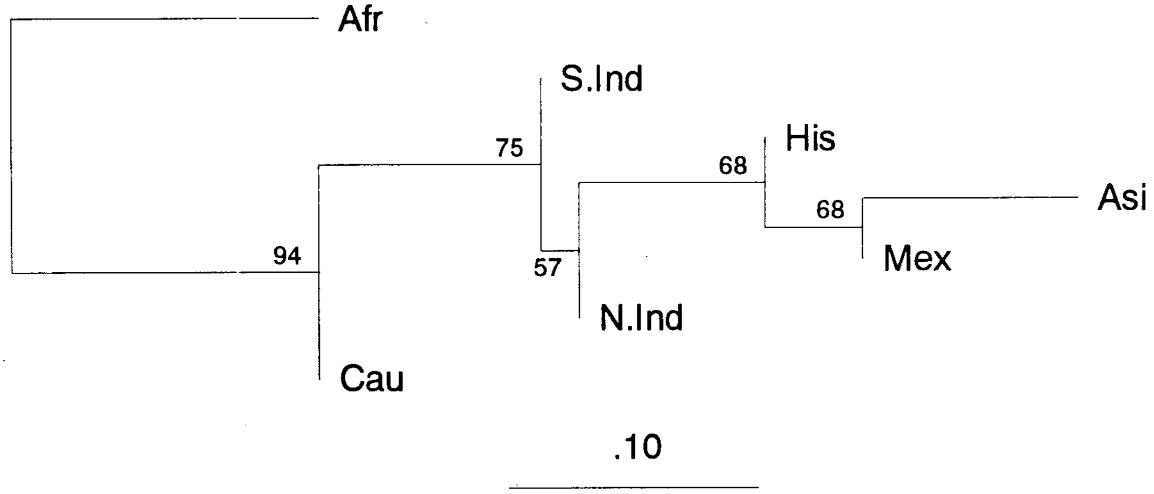

Afr
Mex
68
Asi
68
His
80
98
Ind
Cau
.10

Figure 29: Phylogenetic tree inferred using the Neighbor-Joining algorithm (Saitou and Nei, 1987). Tree was generated by the software package PHYLIP and labels were plotted using the software package TREETOOL.

Phylogenetic tree of seven populations - North Indian (N.Ind), South Indian (S. Ind), African (Afr), Caucasian (Cau), Hispanic (His), Mexican (Mex) and Asian (Asi) using four microsatellite markers. Bootstrap values generated using the CONSENSE program were placed at the nodes.


Afr
S.Ind
75
His
68
Asi
68
94
Mex
57
N.Ind
Cau
.10

was split into the Northern and Southern regions, the same branching pattern was observed for the trees using four and six microsatellite markers. Irrespective of the bootstrap values that were not significant, it was revealed that the North and South Indian populations did not bifurcate from the same point (**see Figures 27 and 29**).

The branch leading to the African population from its closest ancestor was the longest as compared to any other population. Some branches in all trees were negative in length. In all cases, the negative values were small, and branches were assigned a value of zero for length (Saitou and Nei, 1987).

Correlation between Genetic and Geographic Distance of the Four Regions

An analysis of correlation between genetic and geographic distance was computed between the four regions (Northcentral, Northeast, Southeast, Southwest) within India. A correlation coefficient of $r = -0.53$ was revealed. The critical value of r at the 5% level of significance at four degrees of freedom was 0.811. Therefore, the observed absolute value of $r = -0.53$ was less than the critical value of r, therefore, the null hypothesis (H_o: $\rho = 0$) cannot be rejected. **Table 31** shows pairwise genetic and geographic distances between these regions as well as the geographic centre of each region.

Fst

Fst values were computed for six microsatellite loci (CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA) when four populations were pooled (Indian, African, Caucasian, Hispanic) and at four microsatellite loci (CSF1PO, TH01, F13A01, FESFPS) when six populations were pooled (Indian, African, Caucasian, Hispanic, Mexican Asian). When four populations were pooled, Fst values for the six loci ranged between 0.006 and 0.04 with the average over six loci being 0.020. See **table 32** for Fst values. When six populations were pooled, Fst values for the four loci

Table 31: Pairwise Genetic and Geographic Distances between the Four Regions Within India for the Analysis of Correlation.

Correlation Coefficient r = -0.53, df=4, P > 0.05	**Genetic Distance**	**Geographic Distance (km)**
Northeast (Calcutta) - Northcentral (Lucknow)	0.021	950
Northeast (Calcutta) - Southeast (Salem)	0.229	1651
Northeast (Calcutta) - Southwest (Cochin)	0.120	1958
Northcentral (Lucknow) - Southeast (Salem)	0.215	1736
Northcentral (Lucknow) - Southwest (Cochin)	0.138	1954
Southeast (Salem) - Southwest (Cochin)	0.442	320

Table 32: Fst values for Six Microsatellite Loci when Four Populations (Indian, African, Caucasian, Hispanic) are Pooled.

	Loci						
	CSF1PO	**TPOX**	**TH01**	**F13A01**	**FESFPS**	**vWA**	**Average**
Fst	0.006	0.018	0.040	0.033	0.014	0.010	0.020

ranged between 0.004 and 0.047 with the average Fst over four loci being 0.026. See **table 33** for Fst values. A trend was revealed between Fst values of each locus and the percentage of total populations that had allele frequency distributions that were significantly different from that of the Indian population at each locus in the contingency analysis. The TPOX, TH01 and F13A01 loci had Fst values of 0.018, 0.040 and 0.033 respectively. These loci also revealed that 100% of the populations (African, Caucasian, Hispanic) all had significant differences in their allele frequency distributions when compared to the Indian population. The FESFPS, vWA and CSF1PO loci had Fst values of 0.014, 0.010 and 0.006 respectively with the percentage of populations having significant results in the contingency analysis being, 67%, 33% and 33%. The same trend was noted when four microsatellites were used. **Tables 34 and 35** show the results of this trend.

Table 33: Fst values for Four Microsatellite Loci when Six Populations (Indian, African, Caucasian, Hispanic, Mexican, Asian) are Pooled.

	Loci				
	CSF1PO	**TH01**	**F13A01**	**FESFPS**	**Average**
Fst	0.004	0.040	0.047	0.014	0.026

Table 34: Trend between Fst and the Percentage of Populations that have Significant Differences in the Allele Frequency Distributions from that of the Indian Population at Six Microsatellite Loci.

	Loci					
	CSF1PO	**TPOX**	**TH01**	**F13A01**	**FESFPS**	**vWA**
Fst	0.006	0.018	0.040	0.033	0.014	0.010
% of Significant Populations	33	100	100	100	67	33

Table 35: Trend between Fst and the Percentage of Populations that have Significant Differences in the Allele Frequency Distributions from that of the Indian Population at Four Microsatellite Loci.

	Loci			
	CSF1PO	**TH01**	**F13A01**	**FESFPS**
Fst	0.004	0.040	0.047	0.014
% of Significant Populations	20	100	100	40

Discussion

Hardy-Weinberg Equilibrium

Most microsatellite loci are assumed to be selectively neutral since they are primarily found in non-coding regions of DNA. The assumption of selective neutrality was tested for at the six microsatellite markers in the Indian population. A test of Hardy-Weinberg expectations using Monte Carlo simulations for the Indian population revealed significant deviations from Hardy-Weinberg values at the TPOX and vWA loci. Deviations from expected values of Hardy-Weinberg may be due to a variety of causes. If excess heterozygotes are observed in the population, this may be caused by overdominant selection and outbreeding. If excess homozygotes are observed, this may be caused by the presence of null alleles, inbreeding, selection and population structure leading to Wahlund's effect. The link of a microsatellite locus to a disease selected locus would also lead to deviations from Hardy-Weinberg proportions.

The largest division of the Indian population is linguistic. Indo-European languages are spoken in the North and Dravidian languages are spoken predominantly in the South. Natural populations are often found not to be large randomly mating panmictic units. Thus, non-random mating patterns as well as past historical migrations may lead to population structure such as that of the language division within India, thus causing deviations from Hardy-Weinberg expectations.

Relationship between North and South India and other World Populations

Relationship between North and South India

Contingency analysis revealed a significant difference in the allele frequency distribution at the vWA locus between North and South Indian populations. However, the other five markers did not reveal any significant deviations from homogeneity in allele frequency distributions between these two populations. A similar study by Lahermo *et al.*, 1996 used microsatellite and minisatellite markers and contingency analyses to examine the relationship of allele frequency distribution between two subpopulations in Finland. The Finnish and Saami populations have been separated by linguistic and cultural barriers for about a thousand years, much like that of the Aryan(Northern) and Dravidian(Southern) populations of India. However, unlike this study, the Finnish study showed significant differences in the allele frequency distributions between the Finns and Saami using all markers. The lack of significant differences in allele frequency distributions between North and South populations at five of the six loci, may indicate a significant amount of gene flow between these two populations, or it may indicate that these polymorphisms existed before the divergence of the Aryan and Dravidian populations. A recent theory as to why the allele frequency distributions of these populations are so similar, may be due to constraints on the maximum allele size and the high mutation rate at microsatellite loci, thus leading to homogenization effects between populations (Garza *et al.*, 1995).

Previous population studies using microsatellites and minisatellites have shown that in all populations, the allele size range at each locus is essentially the same. In general, the most common alleles were found at very similar size ranges and the main difference seen between populations was the frequency of particular size range of alleles or the frequency of a certain allele (Deka *et al.*, 1995; Deka, 1991; Balazs *et al.*, 1989; Balazs, 1993).

A closer examination of the allele frequency distributions of Northern and Southern populations revealed that at the TPOX and vWA loci, the Southern population contained alleles that were not found in the Northern population. The fact that the Southern population sample size was 1/3 that of the North in this study might suggest that these alleles may be unique to the South or are found in much higher frequencies than the Northern population. At the vWA locus, allele 13 was found in the South Indian population (n=37) with a frequency of 0.026, whereas, this allele was not found in the Northern sample (n=103). Interestingly, this allele was not found in the Caucasian sample (n=212), but was present in the African (n=218) and Hispanic (n=211) populations with frequencies of 0.011 and 0.005 respectively. The higher frequency of this allele in the South Indian population compared to the African and Hispanic population is most probably due to the overestimation in the calculation of allele frequency due to the small sample size. However, the presence of this allele in the South and the lack of this allele in the Northern and Caucasian population, may support the theory that India was originally inhabited by a group of people known as the Dravidians who carried this allele. Later, when the Aryans who lacked this allele entered India, they pushed south the original inhabitants of India. There may be a frequency cline of this allele within India. However, a greater sample size from the South as well as the North Indian populations would be beneficial to get a better measure of the frequency of this allele in the two regions. Furthermore, samples from India's neighboring countries would be useful to see if there is a cline in the frequency of this allele that may reflect past migrational patterns of human populations.

Analysis of the TPOX locus revealed that the South Indian population also carried an allele 13 which was not found in the North Indian sample or any of the other populations. Further investigation of this locus by using a greater sample sizes and more populations as well as pedigree analysis will determine if this allele is the result of a mutation during PCR or if other populations do carry this allele.

Genetic distance calculations between North and South populations displayed different results when a different number of loci were used. The genetic distance between these two populations was not significant when six microsatellite markers were used, however the distance was significant but very small when four markers were used. Small estimates of distance may indicate population substructure in which subpopulations may be randomly mating, but between which there is a reduced amount of gene flow. However, the inconsistency of the results, when four and six markers were used, reflects the need for more loci to be utilized in distance calculations. Furthermore, it also shows that the type of marker or the combination of markers may affect the distance calculations. Therefore, it would be better to have many markers when calculating the genetic distance between two closely related populations.

Phylogenetic analysis using both six and four microsatellite markers failed to show any significant bootstrap values. Thus, little confidence can be placed in the branching order of populations on the tree. However, when both six and four markers were used in the phylogenetic analysis, the North and South Indian populations were not monophyletic. They did not share the most recent common ancestor. This may indicate that the origin of these two populations are different. However, no definite conclusions can be made with respect to the topology of the tree since bootstrap values were not statistically significant.

Previous studies of India have reflected differences between North and South populations since this is the major linguistic division within India. Most of these studies have used gene frequency data from protein or blood group polymorphisms (e.g. Roychoudhury and Nei, 1985; Nei and Roychoudhury, 1993). Phylogenetic analysis has revealed that Dravidian (Southern) populations cluster together, but they do not separate from Indo-European (Northern) speakers at the first or second fission. This indicates considerable admixture between these two groups and although Indo-European speakers are found in the whole subcontinent, they are predominantly found in Northern India while Dravidian speakers occupy the South (Cavalli-Sforza *et al.*, 1994). Further evidence for the theory of language division in India due to past migrations of Dravidians and Aryans has been shown by the recent studies of mitochondrial DNA RFLP's (Restriction Fragment Length Polymorphisms) of the Indian population. A North-South cline in the presence of a particular mitochondrial DNA haplotype has been found. This haplotype is found in higher frequencies in the

South than the North. Furthermore, Caucasian populations seem to have a lower frequency of this haplotype (Passarino *et al.*, 1996). Another study involving the use of shared mitochondrial DNA types by Barnabas *et al.*, 1996 found that the North Indian population appears to have a recent admixture of Caucasian mitochondrial DNA types which were found to be absent in the South Indian population which may indicate that Northern India was populated by a later migration of Caucasian Aryan tribes.

Relationship of North and South Indian Populations to Other Populations

The relationship between the North and South India populations to other world populations was examined. Genetic distance values calculated using six and four microsatellite markers showed that both the North and South Indian populations were closest to Hispanic, Mexican and Asian populations followed by Caucasian, and African populations. However, only the distance values between the Northern and Southern populations and the Caucasian and African populations were significant using both six and four markers. Due to the heterogeneous nature of the Hispanic and Mexican populations, their relationship to the two Indian populations was not addressed in this section, but rather, this problem is addressed in the next section when looking at the Indian population as a whole. Thus, the relationship of the North and South Indian populations to the other populations was examined. The two Indian populations appear to be genetically closer to the Asian population, followed by the Caucasian and African populations. These results are consistent with previous studies in showing that the North and South populations are closer to Caucasian than Africans and they do have some admixture with Asian populations. However, previous studies based on gene frequency data from protein and blood group loci have shown that populations in India are genetically closer to Caucasian populations such as Iranians and Afghans than to East Asian populations (Roychoudhury and Nei, 1985; Nei and Roychoudhury, 1993).

Mitochondrial DNA studies of sequences have shown that Indian populations cluster with Europeans and Chinese connected with relatively short branches thus indicating a recent common ancestor for these groups (Mountain *et al.*, 1995). The close genetic affinity of the North and South populations with Caucasian or West Asian groups has been attributed to gene flow between the Aryan tribes who originated in Southern Russia with the Dravidians who originally inhabited India. Furthermore, a recent study of shared mitochondrial DNA haplotypes by Barnabas *et al.*, (1996) has demonstrated that the South Indian population may have been an earlier Caucasian migration sometime before the more recent arrival of the North Indians. This study showed that, South Indian mitochondrial DNA types shared with Caucasians are placed comparatively closer to the central common DNA type in terms of the number of mutation steps. Whereas, some mitochondrial DNA types shared by Caucasians and North Indians are placed further away from the common mitochondrial DNA type, along the Caucasian branches. Based on these observations, Barnabas *et al.*, (1996) suggests that the population in the South may be an earlier Caucasian or West Asian migration while the North is a mixture with a later West Asian migration. This idea has to some extent been supported by linguistic data. The advent of Indo-European languages in the North could perhaps have coincided with a more recent migration of Caucasians which resulted in the second admixture of Caucasian types of mitochondrial DNA in North India (Barnabas *et al.*, 1996).

As mentioned before, Asian admixture is also present within India, which has been attributed to past invasions (Barnabas *et al.*, 1996). The presence of an ancient East Asian mitochondrial DNA haplotype in high frequencies within the Indian subcontinent and the North-South cline of this haplotype has further added support to the hypothesis that Indians have had some admixture with Asian populations. Passarino *et al.*, (1996) hypothesizes that this haplotype predates the migration of the Indo-European tribes into the subcontinent and most likely predated the split between proto-Indians and proto-Orientals. The arrival of Indo-Europeans who were virtually lacking this haplotype affected population structure mainly in North and in the Center of the subcontinent but not the South, which still carries this haplotype in higher frequencies.

This study using microsatellite markers is consistent with other studies that have demonstrated that North and South Indians are closest to Caucasian populations followed by African. The close relationship between the Indian populations and the Asian population is consistent with earlier studies that used fewer markers, specifically gene frequency data from protein and blood group loci (Roychoudhury, 1977). However, later studies that used more markers have shown that the North and South Indian populations cluster with Caucasian and West Asian populations but do show some admixture with East Asian populations. The genetic distance between the Asian population and the Indian was determined using only four microsatellite markers. Thus, the use of more markers may clarify the relationship between these populations.

Relationship of the Indian Population to Other World Populations

A summary of the contingency analysis revealed a trend between the number of loci that had significant differences in allele frequency distributions when comparing the Indian population to another and the genetic distance between these two populations. A positive correlation was found between these two factors. The greater the number of loci displaying significant differences in allele frequency distributions between the Indian population and another, the greater the genetic distance was between these two populations. The contingency analysis showed that the allele frequency distributions between the Indian and African populations were significantly different at six loci, followed by four for the Caucasian population and three, for the Hispanic population. The genetic distance values between these three populations and the Indian displayed the same relationship. However, only distance values between the Indian and the Caucasian and African populations were significant. When four microsatellite markers were used in the contingency analysis to facilitate the addition of more populations, the same trend was noticed as when six markers were used. The African population's allele frequency distributions were significantly different from that of the Indian population at four loci, followed by three for the Caucasian population and two for the Asian, Hispanic and

Mexican populations. Just as when six markers were used, distances were only significant between the Indian and the Caucasian and African populations. The relationship between the Indian population and the other populations revealed by the contingency and genetic distance analysis is also reflected in the topology of the Neighbor-Joining tree. However, little confidence can be placed in the branching order of the tree due to bootstrap values not being statistically significant. The low bootstrap values reflects the need for a greater number of loci to be used when inferring the phylogenetic relationship between populations (Nei and Takezaki, 1996). Based on the use of four and six microsatellites, it appears that the Indian population is closest to the Asian population followed by the Caucasian and African populations.

Previous studies have demonstrated that the Indian population as a whole clusters more closely with West Asian populations including Turks and Iranians forming a Caucasian or West Asian cluster when compared with Asians from further north and east (Nei and Roychoudhury,1993; Cavalli-Sforza *et al.*, 1994; Roychoudhury, 1983). Phylogenetic studies have also demonstrated that Indian populations reflect some admixture with East Asian populations due to contact in the past (Nei and Roychoudhury, 1993). A study involving shared mitochondrial DNA types demonstrated that the Indian population is closer to Caucasian populations but also has admixture with East Asians (Barnabas *et al.*, 1996). Furthermore, ancient East Asian haplotypes have been found in high frequencies within the Indian subcontinent which may predate the split between Eurasian populations (Passarino *et al.*, 1996). As mentioned in the last section, the close relationship between the Indian population and the Asian is consistent with earlier population genetic studies that used fewer markers (Roychoudhury, 1977). However, the use of six microsatellite markers has shown that Indians are closer to Caucasian than to African populations, and this result is consistent with previous studies based on different markers. The relationship between the Asian and Indian populations may be clarified with the use of more microsatellite loci as was demonstrated in later studies that used a greater number of protein and blood group loci (Nei and Roychoudhury, 1993).

The contingency analysis as well as genetic distance has shown a very close relationship between the Indian and the Hispanic and Mexican populations. The term Hispanic is an ethnic term used to describe an individual of Spanish origin and the term Mexican not only describes the place of origin of a particular individual, but also their ethnicity. For the purpose of this study, these two populations were considered to have the same ethnic background. The Hispanic and Mexican samples were defined by their surname. Definition of an individual by their Spanish surname covers many people of different ethnic backgrounds, including Spanish who are Caucasian and Amerindians who are of East Asian origin. Furthermore, individuals in these samples may be a mix of Spanish and Amerindian. Through mitochondrial DNA studies and gene frequency studies of protein polymorphisms, the origins of the Amerindian tribes who inhabited North, Central and South America are thought to have come from Mongolia (Kolman *et al.*, 1996; Nei and Roychoudhury, 1993). The close relationship of the Indian population to the Hispanic and Mexican populations, may reflect a similar admixture of Caucasian and East Asian elements in these populations. However, it is thought, that the Amerindians went through a severe bottleneck before or during the inhabitation of the Americas. Thus, this could cause the fixation or loss of certain alleles in the population, leading to different allele frequency distributions. Most probably, the Hispanic and Mexican populations have a high admixture of Caucasians which most likely reflects the close relationship with the Indian population. However, with the Hispanic and Mexican populations being heterogeneous in their ethnicity, and without knowing the exact composition of the sample, it is difficult to make any definite statements with regard to the relationship between these two populations and the Indian population. This question may be resolved if populations from Spain and Amerindian tribes that inhabited Central and South America were included in this study as well as the inclusion of more markers.

Relationship of Populations Within India

Pairwise contingency analysis between the Northcentral, Northeast, Southeast and Southwest regions within India revealed only one significant difference in the allele frequency distribution between the Northeast and Southwest populations at the CSF1PO locus. A closer look at the allele frequency distributions revealed that the one characteristic that differed the most between populations was the total numbers of alleles detected at each locus. The number of alleles varied widely between regions. For example, the total number of alleles detected for the CSF1PO locus ranged between 4 and 7. As well, the total number of alleles detected in each region was usually lower than those detected in the entire Indian sample. This most probably reflects the very small sample sizes that were used for each region. Furthermore, small sample sizes inflate allele frequencies, thus contingency analysis results based on allele frequencies may not necessarily reflect the relationships in allele frequency distributions between these regions. Genetic distances computed between these regions were only significant between the Northcentral and Northeast regions and between the Northcentral and Southwest regions. A correlation between genetic and geographic distance revealed a correlation coefficient of -0.53, which was not significant at the 5% level. This indicates no correlation between genetic and geographic distance between these four regions. A number of factors could account for this result, including low sample sizes, not enough microsatellite loci and past migrational movements. Increased sample sizes would give more representative allele frequencies of these regions and the use of more loci would provide more information regarding the relationship between these four regions within India. Population movements in historical times have played an important role in determining the genetic constitution of the Indian subcontinent and previous studies have shown that it is not always possible to identify geographical distance or past ethnic or social connections as the most important determinant of genetic similarity or differentiation within India (Cavalli-Sforza *et al.*, 1994; Papiha *et al.*, 1982). An example of past migrations that may obscure any correlation between genetic and geographic distance has been demonstrated by the Sinhalese of Sri Lanka having a close genetic affinity with people from Northeast India. The Sinhalese are believed to be descendants of people from the regions of Northeast India. Two common traits, hemoglobin E and hairy pinnae are found in both populations and may provide evidence for their contact in the past (Kirk, 1976). Thus, migrational movements in the past may obscure any relationship between genetic and geographic distance within India.

The Use of Microsatellites in Population Genetic Studies

Microsatellites are promising genetic markers for studying the demographic structure and phylogenetic history of populations due to their abundance and high degree of polymorphism. The high rate of molecular evolutionary change makes them particularly useful for studying genetic relationships among closely related populations such as humans. Microsatellites give a more magnified view of recent evolutionary events than do mitochondrial DNA, a difference that can be attributed to their different rates of mutation.

An important assumption in inferring population history from genetic data is that genetic polymorphisms are selectively neutral. Microsatellites occur primarily in non-coding DNA and appear to meet this assumption. Earlier studies used blood groups and protein polymorphisms but these systems may not be selectively neutral and do not directly reflect variation at the DNA level. Mitochondrial DNA has been extensively used in studies of human origins, however, drawbacks of mitochondrial DNA are that it provides a limited amount of information about genetic variation and it may not be selectively neutral.

Heterozygosity

Heterozygosity values for the six loci in the Indian population ranged from 0.69 at the FESFPS locus to 0.82 at the TH01 locus. Observed and expected heterozygosity values were close in the Indian population for each locus. Furthermore comparison of heterozygosity values between the Indian population and the other populations showed them to be close. Comparison of each locus of the Indian population to the African and Caucasian population revealed that observed heterozygosities were generally slightly lower in the Indian population. However, exceptions were seen at the TH01 locus when compared to the African population and the TH01 and TPOX loci when compared to the Caucasian population.

Heterozygosity values for the microsatellite markers used in this study ranged between 0.61 and 0.83 for all six populations. The high values of heterozygosity are consistent with the high rates of mutation found at microsatellite loci. The average observed heterozygosity over six microsatellite loci was the highest in the African population, followed by the Indian, Caucasian and Hispanic populations having very close heterozygosity values. Furthermore, the African population displayed the greatest number of alleles at most loci, compared to any other population.

Previous studies that used microsatellite markers in population genetic studies have also found that heterozygosity values were highest in African populations followed by Caucasian and Asian populations. In addition, the average number of microsatellite alleles per locus was the highest in Africans followed by Europeans, then Asians (Jorde *et al.*, 1995; Bowcock *et al.*, 1994; Nei and Takezaki, 1996). The high diversity of microsatellites among the African populations is in contrast to other nuclear markers such as protein and blood group loci. However, this finding is consistent with data acquired from mitochondrial DNA studies (Cavalli-Sforza *et al.*, 1994). Mitochondrial DNA has been used extensively in studies of human origins and most of these analyses have shown excess genetic diversity in African populations. The high diversity of microsatellites and mitochondrial DNA among African populations has been interpreted as support for the hypothesis of an African origin for humans.

Genetic Distance and Phylogenetic Analysis

A maximum of six microsatellite markers were used to determine the genetic distance between populations as well as to infer the phylogenetic relationship between populations. The distances obtained between populations were significant in some cases and not significant in other cases. Furthermore, when four microsatellite markers were used to calculate genetic distance, distances that were not significant using six markers became significant and vice versa. Cases where distance values became significant with the use of more loci was the distance between the South Indian and Hispanic populations. Cases where distance values became insignificant with the use of more loci was between the Caucasian and Hisapnic populations and between the North and South Indian populations. In the former situation, these distances are more likely to be significant as more loci provide a greater amount of information which better reflects the relationships between populations. In the latter case, the distances that were significant with the use of four loci are most probably not significant. Sampling error due to too few loci may have contributed to a significant result. With regard to the Indian population, only distances between the Indian and Caucasian and African populations were significant. As discussed earlier, the Indian population was genetically closer to the Caucasian population than to the African which is consistent with the results of other studies using different markers.

However, what remained consistent with the use of six and four markers were the distances between the African population and the other populations. They were significant in all cases. Furthermore, the distances between the African population and the others were the largest. However, the distances between the non-African populations were not significant in all situations. This could mean that these populations are very closely related or it may reflect the need for more markers to differentiate between these non-African populations. However, it seems that six microsatellite markers were adequate to differentiate between the African and non-African populations. Phylogenetic analysis revealed bootstrap values that were not significant with the use of six and four microsatellite markers. Thus, little confidence can be placed in the branching order of the tree. However, the node separating the African population from non-African populations was significant or very close to significance in three of the four population trees produced using the Neighbor-Joining method of analysis. However, the internal nodes of the tree did not consistently reveal significant bootstrap values.

Thus, with regard to the use of six microsatellite loci in calculating genetic distance and inferring the phylogenetic relationship between populations, the results suggest that the relationship between African and non-African populations can be determined with confidence, however the relationships between the non-African populations may be made more clear with the use of more microsatellite loci. Furthermore, the results seem to depend on the choice and combination of loci. However, this should not matter with the use of a large number of loci.

Fst

Fst is a measure of the genetic structure of the populations being studied. Fst values were calculated for each locus by pooling four and six populations. A trend was revealed between the results of the contingency analysis and the Fst values. The results showed that the greater the number of populations that had significant differences in allele frequency distributions when compared to the Indian population at a particular locus, the higher the Fst value. Although not all pairwise comparisons were done in the contingency analysis, there was a rough positive correlation between these two factors when only comparisons were done between the Indian population and another. In population studies, markers that have high Fst values show a high amount of variation among populations. This was seen to some extent when pairwise comparisons of allele frequency distributions between the Indian population and other populations were done for each locus. The more populations that were significantly different in allele frequency distributions, the higher the Fst value. The highest Fst values were seen at the TH01 and F13A01 loci, whereby all populations had significant differences in allele frequency distributions from the Indian population. Conversely, the lowest Fst value was at the CSF1PO locus, where only the African population had a significant difference in the allele frequency distribution from the Indian population. Thus, markers with high Fst values are better for differentiating between populations as was demonstrated in this study using microsatellite loci.

Problems with the use of Microsatellites

The mutational processes of microsatellite loci are still being investigated. Thusfar, the Stepwise Mutation Model has been utilized to explain the mutational process of microsatellite repeats. However, this assumption does not hold and there is evidence that mutation sometimes produces nucleotide repeat patterns that require two or more step changes. Therefore, several mathematical models that allow changes of more than one step have been proposed (DiRienzo *et al.*, 1994; Garza *et al.*, 1995). In addition, the irregular pattern of mutation and the limitation in the number of repeats have provided problems in determining which parameters to use in genetic analyses involving microsatellites.

Theoretical mathematical models by Nauta *et al.*, (1996) have demonstrated that high mutation rates combined with limits on the range of allele sizes may lead to the rapid decay of genetic information accumulated by a population. Mutation is generally regarded as a factor that enhances the differentiation between populations. However, when the range of target alleles is limited, and high mutation rates are prevalent, mutation will lead to the reappearance of alleles lost in the past. As a consequence, mutation may be viewed as a homogenizing factor that counteracts the diversifying effects of random genetic drift. Thus, the use of microsatellites to determine genetic divergence may be limited and genetic information specific for a population may easily be lost due to mutation. This idea was proposed as one of the reasons as to why a human population genetics study by DiRienzo *et al.*, (1994) did not show substantial differentiation between the populations using microsatellite loci. Additional explanations offered for these similarities were, selection, gene flow in historic times between these populations or these polymorphisms partially pre-dated the divergence between these populations (Garza *et al.*, 1995). Further investigation of the mutational processes of microsatellite loci as well as additional human population studies are needed to determine if high mutation rates and constraints on allele sizes would affect relationships between closely related populations. However, thusfar, human population genetic studies using microsatellites have shown that phylogenetic reconstruction appears to correlate with results using other types of markers.

The use of microsatellite loci in interspecific phylogeny reconstruction has been less successful since it has been found that microsatellites degrade quickly. These microsatellites are marked by changes in sequence structure that are more complex than changes in repeat number or they are simply not present in one of the species (Rubinsztein *et al.*, 1995; Garza *et al.*, 1995). Garza *et al.*, (1995) has also shown that the average allele sizes at several microsatellite loci in chimpanzees and humans are sufficiently similar, that there may be some constraint on the evolution of average allele size. Therefore, distances at these loci do not appear well correlated with time for more divergent taxa. Garza *et al.*, (1995) also demonstrated that even when microsatellites persist over long periods of time, they fail to reflect separation times among species. Therefore, a potential problem with the use of microsatellite loci is that they will eventually lose their phylogenetic information.

Future Work

Three factors need to be addressed in future human population genetic studies utilizing microsatellite loci. These include population sample sizes, the number of markers and the addition of more populations.

Increasing sample sizes would aid in determining whether certain alleles or groups of alleles are found in higher frequencies or are specific to a particular population. With regard to this study, increasing the South and North Indian samples would determine whether certain alleles are specific to the South or North regions. Also, increasing sample sizes of the four regions within India would clarify questions regarding correlation's between genetic and geographic distances.

The addition of more loci will increase the power of detecting population substructure because each locus will contain an independent history of the population depending on the amounts of random drift, mutation and migration that have occurred. In addition, the loci should be genetically unlinked so results will not be biased towards the events of a single linkage group. Additional loci may clarify relationships between populations within India as well as those between India and other world populations. Measures of genetic distance which are averaged over the total number of loci would better reflect the relationship between these populations if more markers were used and bootstrap values in phylogenetic analyses would become significant.

Additional populations would be useful in determining historical migrational patterns as well as determining the relationship of India to these populations as well as the relationships between these populations. With more populations, especially neighboring populations of India, clines in the frequencies of certain alleles may reflect past migrational patterns. Furthermore, populations that are homogeneous in nature should be used when determining the genetic relationships between populations. Heterogeneous populations with regard to their ethnicity, create further questions when determining the relationships between populations. Finally, the addition of tribal samples from India would also be useful in determining their relationship to the rest of India as well as to each other.

Conclusion

In this study, we attempted to use microsatellite variation in order to differentiate between populations within India that have been divided by spoken language and geographical distance, and between the Indian and other human populations.

Six microsatellite loci that were located in non-coding regions of DNA were typed in a total of 149 individuals of Indian origin.

Differences in allele frequency distributions between North and South Indian populations was detected at one out of six loci. The lack of differences may be attributable to gene flow between these two populations, or that these polymorphisms existed before the split of these two populations. The vWA locus revealed an allele in the South Indian population that was not present in the North Indian or Caucasian population. This may indicate a population specific allele and could lend support to the theory that India was populated by two major groups - Dravidians and Aryans at different times. Furthermore, phylogenetic analyses, although not significant revealed that North and South Indian populations were not monophyletic. However, the results of this study reflect the need for a greater amount of loci and increased sample sizes to detect and confirm any genetic differences between these two populations.

A correlation of genetic and geographic distance between the Northeast, Northcentral, Southeast and Southwest regions of India was not detected. The lack of correlation between these two variables may be due to the small sample sizes of these regions, or past migrational movements which would obscure relationships between genetic and geographic distance.

Genetic distance analyses revealed significant distance values between the Indian, Caucasian and African populations, with the Indian being closer to the Caucasian population.

These results are consistent with earlier studies that used gene frequency and mitochondrial DNA data.

Finally genetic distance and phylogenetic analyses demonstrated that six microsatellite loci are able to distinguish between African and non-African populations, but more loci are needed to differentiate between non-African populations.

Appendix A - Protocols

1. Blood Collection

- Collect blood in 5 ml draw lavender VACUTAINER brand blood collection tubes. Tubes contain 0.05 ml of 15% EDTA (K_3) solution.
- Wear gloves at all times when handling blood and potential DNA samples.
- Place 1 ml aliquots of blood using autoclaved pipet tips into autoclaved1.5 ml microcentrifuge tubes.
- Freeze and store blood aliquots at -70°C.

2. DNA Extraction - Phenol

Whole blood preparation:

- Use 1 ml whole blood in 1.5 ml microcentrifuge tube.
- Spin for 2 minutes at 8000 rpm in a microcentrifuge.
- Remove approximately 0.5 ml of plasma leaving cellular material in tube.

Lyse red blood cells:

- Add 1 ml of cold Red Cell Lysis Buffer (1X).
- Mix by inverting a few times.
- Allow to stand on ice for 15 minutes.
- Spin 2 minutes at 8000 rpm and remove the top layer of liquid. The liquid being removed should be clear (dark red the first few times). Stop where the liquid becomes cloudy. Remove approximately 0.75 ml.

- Resuspend the pellet well and add 1 ml of Red Cell Lysis Buffer. Spin and remove plasma.
- Repeat as necessary (3 lyses should be sufficient).
- When a good white blood cell pellet is obtained, wash with 300 ul of a saline solution and spin at 8000 rpm for 2 minutes. Pipette off the supernatant.

Lyse white blood cells:

- Add 300 ul of SSTE and 100 ul of proteinase K (10 mg/ml): total volume = 400 ul
- Resuspend the pellet.
- Incubate at 55°C for a minimum of 1 hour or until the brown protein pellet has dissolved.
- If suspension is still very gelatinous after incubation, add another 100 ul SSTE and 50 ul of proteinase K.

Separate proteins - Phenol:

- This procedure is done in the fume hood.
- Add an equal volume of phenol to the dissolved protein pellet (350 ul) and mix by inverting.
- Spin for 2 minutes at 8000 rpm.
- Using a pipette, remove the aqueous phase (top layer) and transfer to another autoclaved 1.5 ml epindorf tube. At this point, take any proteins (white stringy material) as well.
- Repeat once more using 350 ul phenol.
- If the aqueous phase is still brown, repeat again using phenol.
- Repeat using a 1:1 mixture of phenol:chloroform iso-amly alcohol (175 ul of each). Avoid the proteins at this point.

- Repeat using 350 ul of chloroform iso-amly alcohol. Stay away from the bottom phase and proteins.

Extract DNA:

- To the last tube, add 1/10th volume of 2M NaAc (approximately 35 ul) and 2-3 times volume of 100% ethanol (approximately 700 ul). Mix by inverting the tube gently a few times.
- Wait until DNA comes out (white strings). When striations are no longer seen between the phases, spin at 14 000 rpm for 6 minutes.
- Pour the ethanol out, being careful not to lose the DNA pellet at the bottom of the tube.
- Add 700 ul of 70% ethanol. Resuspend the pellet and spin at 14 000 rpm for 2 minutes.
- Pour the ethanol out, being careful not to lose the pellet.
- Incubate at 55°C until all the ethanol has evaporated (5-10 minutes).
- Add 50 ul of TE and mix by flicking the tube. Adjust the volume of TE depending on the size of the DNA pellet.
- Incubate at 55°C for 2-3 hours.
- Store at 4°C or freeze at -20°C.

Quantification and dilution of DNA samples:

- Each DNA sample was quantified using a flurometer and the concentration was recorded in ng/ul.
- Aliquots of the concentrated DNA samples were diluted to approximately 20 ng/ul using deionized H_2O for use with the PCR reaction.
- Diluted DNA samples were stored at -20°C.

3. PCR Amplification of 6 Microsatellite Loci

- Two GenePrint™ STR Multiplex kits (Promega) were used to amplify the 6 microsatellite loci followiing the instructions of the manufacturer. One kit contained the loci: CSF1PO, TPOX, TH01 and the other kit contained the loci: FESFPS, F13A01, vWA. The quantities of the reagents added to each PCR reaction were the same for both kits. The only difference was the cycling protocol for the 2 sets of microsatellite loci.
- The use of gloves and sterile procedures is necessary while setting up the PCR reaction to prevent possible contamination.
- Label GeneAmp™ PCR tubes with the sample number. Include a positive and negative control as well.
- In a separate tube, a Master Mix (without template DNA) is made adding 1 or 2 extra reaction volumes to compensate for pipetting error.
- For each sample, including the positive and negative control, a multiplex reaction containing 3 loci requires:

17.35 ul	Sterile deionized water
2.50 ul	STR 10X Buffer (500 mM KCl, 100 mM Tris-HCl, 9.0 at 25°C, 15 mM $MgCl_2$, 1% Triton® X-100, 2 mM of each dNTP)
2.50 ul	Multiplex 10X primer pair mix
0.15 ul	Taq DNA Polymerase (5 Units/ul) (0.75 units)

- Gently mix the Master Mix and place on ice.
- Aliquot 22.5 ul of Master Mix into each reaction tube.
- Add 2.5 ul of template DNA (20 ng/ul) for a total of approximately 50 ng.
- To the negative control, add 2.5 ul of sterile deionized H_2O and to the positive control, add 2.5 ul of K562 DNA (10 ng/ul) for a total of 25 ng.
- The total reaction volume for each tube should be 25 ul.
- To each tube, add 1 drop of mineral oil and close each tube.

- Centrifuge the samples briefly at 14 000 rpm to bring the contents to the bottom of the tube.
- Place in the PCR machine (Perkin Elmer Cetus DNA Thermal Cycler 480) and cycle.
- Cycling protocol #1 to be used with: CSF1PO, TPOX, TH01.

  96°C for 2 minutes, then:

  94°C for 1 minute
  64°C for 1 minute
  70°C for 1.5 minutes
  For 10 cycles, then:

  90°C for 1 minute
  64°C for 1 minute
  70°C for 1.5 minutes
  For 20 cycles, then:

  4°C to soak.
- Cycling protocol #2 to be used with: FESFPS, F13A01, vWA.

  96°C for 2 minutes, then:

  94°C for 1 minute
  60°C for 1 minute
  70°C for 1.5 minutes
  For 10 cycles, then:

  90°C for 1 minute
  60°C for 1 minute
  70°C for 1.5 minutes
  For 20 cycles, then:

  60°C for 30 minutes, then:

  4°C to soak.
- Store PCR products at -20°C.

4. *Agarose Gel Electrophoresis of Amplification Products*

- Wear gloves at all times.

- Prepare a 2% agarose gel by adding 1.0 g of agarose to 50 ml of 1X TAE buffer.
- Heat the agarose solution on a hot plate and bring it to a boil to dissolve the agarose.
- Using the hot agarose solution, seal the sides and corners of the electrophoresis gel box using a Pasteur pipette.
- When the agarose has cooled to about 55°C add approximately 1 ul of ethidium bromide and mix.
- Pour the agarose into the gel tray, insert the comb and let the gel set for 30 minutes.
- When the gel has set, remove the side panels and the comb. Add enough 1X TAE buffer to cover the top of the gel by at least 1 cm.
- Prepare the samples by adding 1 ul of 5X loading solution to 5 ul of PCR product.
- Load 6 ul of sample, positive and negative control onto the gel. Also load 6 ul of a 100 bp ladder.
- Run the gel at approximately 80 volts for 30-45 minutes or until the bands have run half way down the gel.
- Using a UV transilluminator (302 nm), and safety goggles, view and photograph the gel.

5. Polyacrylamide Gel Electrophoresis

- Some of the reagents used in this section were provided by the GenePrint™ STR Multiplex kits and the DNA Silver Staining System kits (Promega). The manufacturer's instructions were followed with some minor modifications.
- It is essential to wear double gloves during this procedure.

Preparation of short glass plate:

- Wear gloves at all times. This procedure is done under the fume hood.

- Using only a scrupulously clean plate, wipe it with a KimWipe™ saturated with freshly prepared binding solution (1 ml). Make sure the plate is completely covered.
- Let it dry for 4-5 minutes.
- Apply approximately 2 ml of 95% ethanol to the plate and wipe with a paper towel in one direction and then perpendicular to the first direction using gentle pressure. Rubbing hard will remove an excessive amount of the binding solution and the gel may not adhere as well.
- Repeat this wash once more.
- It is essential to prevent the binding solution from contaminating the long glass plate. Therefore, different utensils should be used for each plate to prevent contamination.

Preparation of long glass plate:

- Wear new gloves. This procedure is done under the fume hood.
- On a scrupulously clean plate, pour about 2-3 ml of SigmaCote® solution. Spread evenly with a paper towel.
- Let dry for about 5-10 minutes.
- Remove any excess SigmaCote® using a KimWipe® tissue. Excess SigmaCote® may cause inhibition of staining.

To make mould:

- Rinse the spacers (0.4 mm) and the 2 shark tooth combs (0.4 mm) with ethanol, making sure that they are clean.
- Place the spacers on the long glass plate and carefully place the short glass plate over top making sure not to touch the long glass plate.
- Clamp one side of the plates to prevent movement and fit the rubber clamp around the 2 plates in order to hold them together (taping is acceptable).

Preparing a 6% denaturing acrylamide solution:

- In an Erlenmeyer flask, combine the following reagents:

Urea	31.50 g	7 M
Deionized H2O	36.25 ml	
10X TBE	3.75 ml	0.5X
40% acrylamide:bis (19:1)	11.25 ml	6%
Total Volume	**75 ml**	

- Filter the solution using a vacuum filter.
- Store the 6% acrylamide solution in dark bottles at 4°C.

Pouring the gel:

- To 75 ml of 6% acrylamide solution, add 500 ul of 10 % APS and 50 ul of TEMED and swirl.
- Pour the gel using a 75 ml syringe (no needle). Hold the mould at a 45° angle and pour the acrylamide along one edge of the plate.
- After pouring, lay the gel at a 10° angle.
- Insert the 2 combs between the glass plates (about 6 mm). Secure the combs with 2-3 clamps.
- Allow the gel to polymerize for 1.5 hours.
- The gel may be stored overnight by saturating paper towels with 0.5X TBE and wrapping the end with the combs. Plastic wrap is then placed around the wells to prevent the paper towels from drying out.

Gel pre-run:

- Remove the rubber clamp from the polymerized acrylamide gel and clean the glass plates with paper towels and deionized water.
- Remove the comb and any excess acrylamide.
- Secure the mould (glass plates) to the sequencing gel apparatus.
- Add 0.5X TBE buffer to the top (500 ml) and bottom chambers (500 ml).

- Using a syringe, clean out the well, making sure that all of the polymerized acrylamide gel pieces have been removed. Remove all bubbles between the well also.
- The gel is about 40 cm long. Therefore, pre-run the gel at 2000 Volts, 65 Watts and 65 mAmps for 45-60 minutes.

Sample preparation:

- Prepare PCR samples by mixing 2.5 ul of each sample with 2.5 ul of STR 2X Loading Solution.
- Prepare the positive control, pGEM® DNA markers and the STR Ladder mix by adding 2.5 ul of each to 2.5 ul of 2X STR Loading Solution.
- pGEM® DNA markers are run at either end of the gel and positive controls and STR adders are run every 3-4 samples.
- Briefly spin the tubes containing the samples, positive controls, STR ladders and pGEM® DNA markers to bring the contents to the bottom.
- Store the samples at 4°C or at -20°C until ready to load.

Sample loading and gel electrophoresis:

- After the pre-run of the gel, use a syringe to flush the well to clean out the urea and acrylamide pieces. Insert the shark tooth combs with the teeth penetrating the gel about 1-2 mm. The combs are left in the gel during gel loading and electrophoresis.
- Denature the samples by heating at 95°C for 2 minutes and immediately chill on ice.
- Load 2.5 ul of sample into the respective wells. The loading process should take no longer than 20 minutes to prevent the gel from cooling.
- The gel is run at 2000 Volts, 65 Watts, 65 mAmps for approximately 1.5 hours or until the second blue dye (xylene cyanol) is about 15 cm from the bottom of the gel.

6. Silver Staining

- Silver staining was done using the DNA Silver Staining System kits (Promega). Some modifications were made to the manufacturer's instructions.
- After electrophoresis, carefully separate the plates using a wedge. The gel should be strongly affixed to the short glass plate.
- A minimum of 2 Nalgene® tubs are required for this procedure and gloves must be worn at all times.
- Place the plate with the gel in a Nalgene® plastic tub and add 2000 ml of fix/stop solution, making sure that the gel is covered.
- Agitate the plate for 20 minutes until the tracking dye (xylene cyanol) is no longer visible. Save the fix/stop solution to terminate the developing reaction (later).
- Prepare the developer solution without the formaldehyde and the sodium thiosulfate and chill to between 4-10°C.
- Rinse the gel 3 times (approximately 2 minutes) with ultrapure water (about 1300 ml) using agitation. Lift the gel (plate) out of the wash and allow it to drain 10-20 seconds before transferring it to the next wash.
- Transfer the gel to a tub containing 2000 ml of staining solution and agitate for 30 minutes.
- Complete the preparation of the developer solution 5 minutes before the staining procedure is finished.
- Remove the gel from the staining solution and dip it briefly into a tub containing ultrapure water (1300 ml) then place the gel in the developer solution (1000 ml).
- Agitate the gel for 2-4 minutes or until the template bands start to develop and faint bands are visible.

- Transfer the gel to fresh developer solution (1000 ml) and agitate for and additional 2-3 minutes until all bands become visible.
- Stop the developing reaction by adding 1000 ml of fix/stop solution and agitate for 2-3 minutes. It is better to stop the developing reaction early to prevent a high background on the gel from occurring.
- Rinse the gel twice for 2 minutes each with approximately 1300 ml of ultrapure water.
- Dry the gel overnight by standing it upright at room temperature.

7. APC Film Development

- In a darkroom, with the safelight on, place the dry, stained gel attached to the plate (gel side up) on a white fluorescent light box. The gel must be completely dry.
- Position the APC Film, emulsion side down, over the gel to be copied. The emulsion side has a glossy white surface.
- Turn on the light box and expose the film for 30-60 seconds depending on how dark or light the gel has been stained.
- Develop the film using an automatic film processor.
- The APC film produces a direct positive, mirror image of the original gel.
- Alleles are scored either directly from the gel or from the film.
- The glass plates can be re-used by placing them in a solution of 10% sodium hydroxide for 1 hour to overnight. The sodium hydroxide solution removes the gel and the SigmaCote®.

Appendix B - Solutions

Stock Solutions

EDTA

Concentration:	0.5 M
pH:	7.4 & 8.0
Amount:	1 L
Autoclave:	Yes

186.12 g	$Na_2EDTA.H_2O$	**M.W.** = 372.2 g/mol
800 ml	DDW (double distilled H_2O)	

- adjust pH with NaOH pellets
- adjust volume to 1 L with DDW

HCL

Concentration:	1 M
pH:	
Amount:	500 ml
Autoclave:	No

18.23 g	HCL	**M.W.** = 36.46 g/mol
400 ml	DDW	

NaOH

Concentration:	1 M
pH:	
Amount:	500 ml
Autoclave:	No

20.00 g	NaOH	**M.W.** = 40.00 g/mol
400 ml	DDW	

- adjust volume to 500 ml with DDW

NaCl

Concentration:	5 M
pH:	
Amount:	500 ml

146.1 g	NaCl	**M.W.** = 58.44 g/mol
400 ml	DDW	

- adjust volume to 500 ml with DDW

Concentration: 6 M
pH:
Amount: 100 ml
Autoclave: Yes

35.07 g	NaCl	**M.W.** = 58.44 g/mol
80 ml	DDW	

- adjust volume to 100 ml with DDW

SDS

Concentration: 10%
pH:
Amount: 500 ml
Autoclave: Yes

50.00 g	SDS	**M.W.** = n/a
400 ml	DDW	

- adjust volume to 500 ml with DDW

Concentration: 20%
pH:
Amount: 100 ml
Autoclave: Yes

20.00 g	SDS	**M.W.** = n/a
70 ml	DDW	

- adjust volume to 100 ml with DDW

STE

Concentration: 1X
pH:
Amount: 500 ml
Autoclave: Yes

10.0 ml	5 M NaCl	0.1 M
12.5 ml	2 M Tris (pH 7.5)	0.05 M
1.0 ml	0.5 M EDTA (pH 7.4)	0.001 M
450 ml	DDW	

- adjust pH with NaOH pellets

- adjust volume to 500 ml with DDW

Tris (Tris-HCl)

Concentration:	2 M
pH:	7.5 & 8.0
Amount:	1 L
Autoclave:	Yes

242.3	g	Tris	**M.W.** = 121.14
800	ml	DDW	

- adjust pH with HCl
- adjust volume to 1 L with DDW

Buffers

TAE (Tris Acetate)

Concentration:	50X
pH:	7.2
Amount:	1 L
Autoclave:	No

242	g	Tris base
57.1	ml	Glacial acetic acid
100	ml	0.5 M EDTA @ pH 8.0
700	ml	DDW

- adjust volume to 1 L with DDW
- dilute to 1X using DDW

TBE (Tris Borate)

Concentration:	10X
pH:	8.3
Amount:	1 L
Autoclave:	No

107.8	g	Tris base
7.44	g	EDTA
55.0	g	Boric acid
700	ml	DDW

- adjust the pH slowly with boric acid pellets
- adjust the volume to 1 L with DDW

DNA Extractions

Red Cell Lysis Buffer (NH_4Cl Na_4HCO_3)

Concentration:	10X
pH:	
Amount:	1 L
Autoclave:	Yes

70.0	g	NH4Cl	1.31 M	**M.W.** = 53.45 g/mol
0.71	g	NH4HCO3	0.009 M	**M.W.** = 78.98 g/mol
800	ml	DDW		

- adjust volume to 1 L with DDW
- dilute to 1X with DDW and autoclave

Saline

Concentration:	0.9%
pH:	
Amount:	500 ml
Autoclave:	Yes

4.5	g	NaCl
400	ml	DDW

- adjust volume to 500 ml with DDW

SSTE (white blood cell lysis buffer)

Concentration:	0.5% SDS in STE
pH:	
Amount:	100 ml
Autoclave:	Yes

5	ml	10% SDS
95	ml	STE

NaAc (sodium acetate)

Concentration:	2 M
pH:	
Amount:	100 ml
Autoclave:	Yes

27.22	g	NaAc	**M.W.** = 136.1 g/mol
80	ml	DDW	

- adjust volume to 100 ml with DDW

TE

Concentration:	1X
pH:	7.7
Amount:	500 ml
Autoclave:	Yes

2.5	ml	Tris @ pH 7.5	10 mM
1.0	ml	EDTA @ pH 8.0	1 mM
450	ml	DDW	

- adjust pH with HCl
- adjust volume to 500 ml with DDW

Proteinase K

Concentration:	10 mg/ml
pH:	
Amount:	
Autoclave:	No

- add 10 ml DDW to 100 mg of Proteinase K. Aliquot into autoclaved 1.5 ml epindorf tubes and store at -20°C.

Ethanol (EtOH)

Concentration:	95% & 70%
pH:	
Amount:	
Autoclave:	No

PCR Amplification

Kit:	GenePrint™ STR Multiplex Systems
Loci:	CSF1PO, TPOX, TH01 and F13A01, FESFPS, vWA
Manufacturer:	Promega

Reagents supplied by the kit:

10X STR Buffer	1.2 ml	500 mM Kcl 100 mM Tris-Hcl, pH 9.0 @ 25°C 15 mM $MgCl_2$ 1% Triton® X-100 2 mM each dNTP
10X Primer Pairs	250 ul	(5 uM)
K562 HMW DNA	3 ug	(10 ng/ul)

- store all reagents @ -20°C

- the rest of the reagents supplied by these kits are used during polyacrylamide gel electrophoresis

Other reagents needed:

Deionized water

Mineral oil (Sigma Chemical Co.)

Taq DNA Polymerase (5 units/ul) (Boehringer Mannhiem)

Agarose Gel Electrophoresis

Agarose Gel

Concentration:	2%
pH:	
Amount:	50 ml
Autoclave:	No

1.0	g	agarose
50.0	ml	1X TAE

- bring to a boil and let cool to 50 °C before pouring
- add 1 ul of ethidium bromide before pouring

Loading Solution

Concentration:	5X
pH:	
Amount:	
Autoclave:	No

5%		Ficoll® 400
0.1%		bromophenol blue
0.1 %		xylene cyanol
100	mM	EDTA pH 8.0
10	mM	Tris-HCl pH 7.5

Polyacrylamide Gel Electrophoresis

Kit:	GenePrint™ STR Multiplex Systems
Loci:	CSF1PO, TPOX, TH01 and F13A01, FESFPS, vWA
Manufacturer:	Promega

Reagents supplied by the kits:

STR Ladder Mix	125 ul
STR 2X Loading Solution	1 ml
pGEM® DNA	3 ug (20 ng/ul)

Other reagents needed:

Kit:	DNA Silver Staining System
Manufacturer:	Promega

Reagents supplied by the kit:

Bind Silane	500 ul

- store @ 4°C
- the rest of the reagents provided by this kit is used in the silver staining procedure

Other reagents needed:

Acetic Acid in Ethanol

Concentration:	0.5%
pH:	
Amount:	200 ml
Autoclave:	No

1	ml	Glacial acetic acid
199	ml	95 % ethanol

Acrylamide:bis (19:1)

Concentration:	40%
pH:	
Amount:	1 L
Autoclave:	No

380	g	Acrylamide
20	g	Bisacrylamide
500	ml	DDW

- adjust volume to 1 L with DDW

Acrylamide Solution (6%)

Concentration:	6%
pH:	
Amount:	75 ml
Autoclave:	No

31.50	g	Urea	7 M
3.75	ml	10X TBE pH 8.3	0.5X
11.25	ml	40% acrylamide:bis (19:1)	6%
36.25	ml	DDW	

- adjust volume to 75 ml with DDW
- filter the solution (vacuum)

Ammonium Persulfate

Concentration:	10%
pH:	
Amount:	5 ml
Autoclave:	No

0.5	g	ammonium persulfate
5	ml	DDW

- store in 500 ul aliquots at -20 °C

TEMED (GIBCO- BRL)

SigmaCote®

0.5X TBE pH 8

Silver Staining

Kit:	DNA Silver Staining System
Manufacturer:	Promega

Reagents supplied by the kit:

Silver Nitrate	20 g
Formaldehyde (37%)	60 ml
Sodium Thiosulfate (10 mg/ml)	10 ml
Sodium Carbonate	600 g

Fix/Stop Solution

Concentration:	10%
pH:	
Amount:	2 L
Autoclave:	No

200	ml	Glacial acetic acid
1800	ml	DDW

Staining Solution

Concentration:	
pH:	
Amount:	2 L
Autoclave:	No

2	g	$AgNO_3$
3	ml	Formaldehyde (37%)
2	L	DDW

Developing Solution

Concentration:	
pH:	
Amount:	2 L
Autoclave:	No

60	g	Na_2CO_3
3	ml	Formaldehyde (37%)
400	ul	Sodium thiosulfate (10 mg/ml)
2	L	DDW

- add the sodium carbonate to 2 L of DDW and chill to between 4-10 °C before adding the formaldehyde and the sodium thiosulfate.

Appendix C - Data

Allele Frequencies for Six Loci in the Indian Population

CSF1PO	Indian		CSF1PO	N. Indian		CSF1PO	S. Indian	
n = 148	N	AF	n = 111	N	AF	n = 37	N	AF
6	0	0	**6**	0	0	**6**	0	0
7	0	0	**7**	0	0	**7**	0	0
8	1	0.003	**8**	1	0.005	**8**	0	0
9	11	0.037	**9**	10	0.045	**9**	1	0.013
10	69	0.233	**10**	55	0.25	**10**	14	0.184
11	92	0.312	**11**	61	0.272	**11**	31	0.421
12	99	0.334	**12**	78	0.35	**12**	21	0.289
13	19	0.064	**13**	13	0.059	**13**	6	0.079
14	4	0.014	**14**	3	0.014	**14**	1	0.013
15	1	0.003	**15**	1	0.005	**15**	0	0

TPOX	Indian		TPOX	N. Indian		TPOX	S. Indian	
n = 148	N	AF	n = 111	N	AF	n = 37	N	AF
6	0	0	**6**	0	0	**6**	0	0
7	0	0	**7**	0	0	**7**	0	0
8	119	0.402	**8**	93	0.423	**8**	26	0.342
9	39	0.132	**9**	33	0.145	**9**	6	0.092
10	26	0.088	**10**	18	0.082	**10**	8	0.105
11	102	0.345	**11**	71	0.318	**11**	31	0.421
12	9	0.030	**12**	7	0.032	**12**	2	0.026
13	1	0.003	**13**	0	0	**13**	1	0.013

TH01	Indian		TH01	N. Indian		TH01	S. Indian	
n = 148	N	AF	n = 111	N	AF	n = 37	N	AF
5	0	0	**5**	0	0	**5**	0	0
6	83	0.280	**6**	63	0.281	**6**	20	0.276
7	49	0.166	**7**	40	0.182	**7**	9	0.118
8	43	0.145	**8**	30	0.132	**8**	13	0.184
9	79	0.267	**9**	58	0.264	**9**	21	0.276
9.3	40	0.135	**9.3**	30	0.136	**9.3**	10	0.132
10	2	0.007	**10**	1	0.005	**10**	1	0.013
11	0	0	**11**	0	0	**11**	0	0
12	0	0	**12**	0	0	**12**	0	0

F13A01 Indian			F13A01 N. Indian			F13A01 S. Indian		
n = 125	N	AF	n = 95	N	AF	n = 30	N	AF
3.2	36	0.144	3.2	25	0.133	3.2	11	0.177
4	14	0.056	4	12	0.064	4	2	0.032
5	96	0.384	5	69	0.356	5	27	0.468
6	45	0.180	6	37	0.197	6	8	0.129
7	41	0.164	7	35	0.186	7	6	0.097
8	0	0	8	0	0	8	0	0
9	0	0	9	0	0	9	0	0
10	0	0	10	0	0	10	0	0
11	0	0	11	0	0	11	0	0
12	1	0.004	12	1	0.005	12	0	0
13	1	0.004	13	0	0	13	1	0.016
14	4	0.016	14	2	0.011	14	2	0.032
15	4	0.016	15	2	0.011	15	2	0.032
16	8	0.032	16	7	0.037	16	1	0.016

FESFPS Indian			FESFPS N. Indian			FESFPS S. Indian		
n = 124	N	AF	n = 94	N	AF	n = 30	N	AF
7	0	0	7	0	0	7	0	0
8	0	0	8	0	0	8	0	0
9	1	0.004	9	1	0.005	9	0	0
10	41	0.165	10	29	0.151	10	12	0.210
11	109	0.44	11	85	0.457	11	24	0.387
12	61	0.246	12	45	0.242	12	16	0.258
13	36	0.145	13	28	0.145	13	8	0.145
14	0	0	14	0	0	14	0	0

vWA Indian			vWA N. Indian			S. Indian		
n = 140	N	AF	n = 103	N	AF	n = 37	N	AF
11	0	0	11	0	0	11	0	0
12	0	0	12	0	0	12	0	0
13	2	0.007	13	0	0	13	2	0.026
14	28	0.100	14	18	0.088	14	10	0.131
15	17	0.061	15	11	0.054	15	6	0.079
16	65	0.232	16	44	0.216	16	21	0.276
17	87	0.311	17	70	0.343	17	17	0.224
18	55	0.196	18	44	0.206	18	11	0.171
19	23	0.082	19	18	0.088	19	5	0.066
20	3	0.011	20	1	0.005	20	2	0.026
21	0	0	21	0	0	21	0	0

Genotypes of Samples of Indian Origin

	CSF1PO	TPOX	TH01	F13A01	FESFPS	vWF
SAMPLE						
1	12/11	11/8	7/6	5/5	12/11	17/17
2	11/11	11/8	8/6	7/4	12/11	16/14
3	11/10	12/8	9/6			17/17
4	12/10	11/11	9.3/9	5/5	12/11	17/17
5	13/11	11/8	9.3/7	6/5	12/12	18/14
6	12/11	11/9	8/6	5/5	13/10	18/18
7	12/11	11/8	7/6			18/16
8	13/12	9/8	7/6	7/5	13/11	18/17
9	15/11	8/8	9/7			
10	12/11	8/8	9.3/9		12/10	19/16
11	11/9	11/8	9/6	7/5	12/10	17/14
12	12/11	11/10	9.3/7	7/5	11/11	19/18
13	12/10	12/8	9.3/9.3	16/14	12/12	18/17
14	12/11	9/9	9/7	6/5	13/11	16/14
15	11/10	8/8	9/9	5/3.2		18/17
16						
17	11/10	8/8	9.3/6			18/14
18	12/11	8/8	8/7	16/3.2	13/13	17/15
19				5/3.2		17/17
20	12/10	10/9	8/6	3.2/3.2		19/17
21						
22	13/11	9/9	6/6			
23	12/12	11/10	9/6	12/5	11/10	17/14
24	11/10	11/8	9.3/6	7/5	13/11	18/17
25	13/11	12/8	9/6	5/3.2	11/10	18/17
26						
27	12/10	9/8	9.3/9	15/5	13/12	16/16
28	12/10	12/8	6/6	16/5	13/11	18/14
29						
30	12/11	8/8	9/8	7/5	13/11	18/16
31	10/10	11/9	9.3/7	6/3.2	11/11	16/16
32	12/11	8/8	8/7			20/16
33	12/9	11/11	9/6			19/17
34	12/12	9/8	9.3/7	6/6	13/13	19/14
35	12/10	10/8	9.3/9	5/5	11/11	18/18
36	14/12	11/10	9/6	15/14	10/10	19/17
37	11/11	10/8	9.3/9		11/11	19/18
38	11/9	8/8	6/6			
39	12/11	11/11	9.3/6			16/15
40	12/11	8/8	7/6	6/4	12/11	17/17
41	12/11	11/8	9.3/6	5/3.2	11/11	17/15
42	13/13	11/10	9.3/7			16/15

43	11/10	10/9	9/8	5/5	13/12	16/14
44	12/10	12/10	7/6	5/3.2	12/11	16/16
45	10/10	8/8	6/6	5/3.2	11/10	17/14
46	12/10	8/8	9.3/9.3	6/4	11/10	16/16
47	11/10	11/9	9/9	5/5	11/10	18/17
48	12/11	8/8	9/6	3.2/3.2	11/10	18/17
49	11/10	11/10	9/7	6/6	13/13	15/14
50	12/9	10/8	8/6	7/5	13/11	18/18
51	11/10	11/8	6/6	5/5	13/10	19/15
52	12/10	11/8	9.3/9			17/14
53	11/10	11/10	6/6	6/6	12/11	19/16
54	11/11	13/12	8/6	5/3.2	12/11	17/16
55	12/12	11/11	9/8	5/5	12/12	14/14
56	10/9	11/8	9/7	6/5	11/11	17/16
57	13/11	9/8	6/6	7/5	12/11	18/17
58	12/12	11/8	9/8			18/15
59	11/10	11/9	9/6	7/5	12/11	20/14
60	10/10	10/8	8/6			15/14
61	12/10	9/8	7/7	7/6	11/11	17/14
62	12/11	9/8	9/6	5/3.2	11/11	17/15
63	12/10	8/8	9/9			
64	11/11	10/8	8/8	5/5	13/10	18/14
65	13/12	11/11	8/6	6/5	12/11	17/16
66	14/10	11/9	10/9	5/3.2	12/11	19/17
67	11/10	10/8	9.3/7	6/3.2	11/11	18/17
68	12/11	11/9	9/7	15/6	11/10	18/18
69	12/10	11/8	9/6	7/5	12/11	18/17
70	12/12	11/8	9.3/9	16/7	11/11	19/17
71	10/8	11/8	9/7			
72	11/11	11/8	9/8	6/5	12/10	17/16
73	12/10	11/9	8/7	7/5	12/11	17/16
74	10/10	11/9	9/7	16/6	11/10	17/16
75	11/10	8/8	7/6	7/3.2	10/10	18/16
76	12/9	11/8	9/6	7/4	12/11	17/16
77	12/11	10/8	9.3/6	7/5	11/10	17/16
78	11/10	11/11	9.3/9			
79	13/11	11/8	9/7	4/4	12/11	15/14
80	10/10	8/8	9/7	7/4	12/11	17/17
81	11/11	11/9	6/6	7/5	11/11	18/17
82	11/10	11/8	9/6	6/3.2	12/10	18/16
83	13/11	11/8	9/8	5/5	12/11	16/16
84	11/11	11/8	9/6	5/5	11/11	16/14
85	12/11	11/9	9.3/7	6/5	13/11	18/16
86	12/12	11/10	8/6	6/3.2	12/12	17/17
87	11/11	12/9	9/6	7/3.2	12/11	17/17
88	12/9	9/8	9/7	5/5	12/11	18/17
89	12/11	11/10	8/6	7/5	11/11	17/17

90	11/10	11/11	9.3/9	6/3.2	12/11	19/14
91	12/11	8/8	9/6	7/5	12/12	18/18
92	11/10	11/8	8/6	7/6	11/11	19/15
93	12/10	11/9	9.3/8			
94	12/12	8/8	9.3/9	3.2/3.2	13/11	17/16
95	13/13	11/11	9.3/9	6/5	12/11	19/16
96	11/10	9/8	8/8	7/6	11/11	18/17
97	12/10	11/11	9/6	16/6	11/11	17/17
98	12/10	10/8	8/7			20/17
99	12/12	9/8	7/6	6/5	13/11	17/15
100	11/9	9/8	9/6	6/4	11/11	17/16
101	11/10	11/8	9/6	5/3.2	11/11	18/16
102	12/11	11/10	9/9	7/6	13/11	18/16
103	12/12	8/8	8/6	7/6	11/11	18/17
104	12/12	8/8	9/8	5/5	11/10	17/14
105	12/10	9/8	9/6	6/5	13/10	18/14
106	10/10	10/8	9.3/6	5/5	13/13	17/17
107	11/11	10/9	8/7	6/5	12/12	18/17
108	12/12	11/11	7/6	14/7	13/11	17/14
109	12/12	11/8	9.3/9.3	6/5	11/10	19/15
110	12/10	12/11	7/6	6/5	12/11	19/18
111	11/10	11/8	9.3/9	5/3.2	12/11	18/16
112	12/11	10/9	9/6	7/5	13/12	17/17
113	13/12	11/8	9/7	16/7	11/10	19/19
114	13/12	11/10	8/6	4/3.2	13/10	16/16
115	13/12	11/11	8/7	7/4	13/12	16/14
116	11/9	11/8	9/8	5/4	12/11	18/17
117	12/12	9/8	9/6	5/5	13/12	17/16
118	14/10	9/8	9/6	7/5	11/10	16/16
119	12/10	11/8	7/6			
120	12/10	11/11	8/7	7/6	11/11	18/18
121	11/10	11/8	9/9	6/5	13/11	18/18
122	12/9	11/8	9/9	5/3.2	11/10	16/15
123	12/11	9/8	8/6	6/3.2	13/13	15/15
124	12/12	11/10	7/7	7/3.2	12/10	17/16
125	12/11	10/8	6/6	6/6	13/12	16/14
126	11/10	9/8	9/7	15/5	13/12	19/17
127	10/10	11/8	10/6	7/5	12/11	13/13
128	12/9	11/8	8/8	3.2/3.2	11/11	18/17
129	11/11	11/9	9/9			19/16
130	13/10	11/8	7/6	7/5	11/10	18/17
131	12/11	11/8	9/8	5/5	11/10	17/16
132	12/12	11/8	9.3/8	5/4	11/10	16/14
133	11/11	8/8	9/7	5/5	12/12	17/16
134	12/10	11/11	9.3/8	5/5	10/10	18/16
135	11/10	11/11	9/9	7/5	12/11	17/16
136	12/11	11/11	9/6	5/3.2	12/11	16/16

137	12/10	11/11	9.3/6	16/14	12/10	17/16
138	13/10	11/8	8/6			17/16
139	11/11	8/8	8/7	7/5	11/10	19/16
140	13/11	11/8	8/7	5/5	12/11	18/17
141	14/10	10/8	7/6	7/6	12/10	17/15
142	11/10	12/8	9.3/7	6/5	11/10	17/16
143	12/11	8/8	9.3/6	7/5	11/10	17/16
144	12/11	9/8	9/6	4/3.2	12/12	17/16
145	10/10	11/8	8/6	6/3.2	12/9	18/17
146	11/11	11/11	9.3/6	5/5	11/11	17/16
147	12/11	9/8	7/6	6/6	12/11	18/17
148	12/11	9/8	9/8	13/4	11/10	16/14
149	12/12	11/8	9/6	7/3.2	13/12	19/17
150	12/11	11/8	9/7	7/3.2	11/10	19/16
151	13/11	11/11	6/6	6/5	11/11	18/16
152	12/10	8/8	9.3/7			
153	12/12	11/11	9.3/9	7/5	13/10	18/17

References

Ashley M.V., and Dow B. D. (1994) The use of microsatellite analysis in population biology: Background, methods and potential applications. *in*: " *Molecular Ecology and Evolution*: *Approaches and Applications*," Birkhauser Verlag Basel, Switzerland.

Balazs I., Neuweiller J., Gunn P., Kidd J., Kidd K.K., Kuhl J., and Mingjun L. (1992) Human population genetic studies using hypervariable loci. Analysis of Assamese, Australian, Cambodian, Caucasian, Chinese and Melanesian populatons. *Genetics*. **131**: 191-198.

Balazs I., Baird M., Clyne M., and Meade E. (1989) Human population genetic studies of five hypervariable DNA loci. *American Journal of Human Genetics*. **44**: 182-190.

Balazs I. (1993) Population genetics of 14 ethnic groups using phenotypic data from VNTR loci. *in*: "*DNA Fingerprinting: State of the Science*," Birkhauser Verlag Basel, Switzerland.

Barnabas S., Apte R.V., and Suresh C.G. (1996) Ancestry and interrelationships of the Indians and their relationship with other world populations: A study based on mitochondrial DNA polymorphisms. *Annals of Human Genetics*. **60**: 409-422.

Begley V. *Encyclopedia Americana. Deluxe Library Edition*. Grolier Inc., Dunbury, **Vol. 14**. 1994.

Behara A.M.P. (1995) *Mitochondrial DNA sequence variation and human population structure in the Indian subcontinent* (M.Sc. Thesis) McMaster University.

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

Bruford M.W., and Wayne R.K. (1993) Microsatellites and their application to population genetic studies. *Current Opinion in Genetics and Development*. **3**: 939-943.

Budowle B., Giusti A.M., Waye J.S., Baechtel F.S., Fourney R.M., Adams D.E., Presley L.A., Deadman H.A., and Monson K.L. (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. *American Journal of Human Genetics*. **48**: 841-855.

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

Cavalli-Sforza L.L., Menozzi P., and Piazza A. (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton New Jersey.

Chakraborty R., Walter H., Mukherjee B.N., Malhorta K.C., Sauber P., Banerjee S., and Roy M. (1986) Gene differentiation amoung ten endogamous groups of West Bengal, India. *American Journal of Physical Anthropology.* **71**: 295-309.

Cutler N. Dravidian Languages and Literatures. *Encyclopedia of Asian History.* **Vol. 1.** 1988.

Das B.M., Das P.B., Das R., Walter H., and Danker-Hopfe H. (1986) Anthropological studies in Assam, India. *Anthrop. Anz.* **3**: 239-248.

Deka R., Chakraborty R., and Ferrel R.E. (1991) A population genetic study of six VNTR loci in three ethnically defined populations. *Genomics.* **11**: 83-92.

Deka R., Jin L., Shriver M.D., Yu L.M., DeCroo S., Hundrieser J., Bunker C.H., Ferrel R.E., and Chakraborty R. (1995) Population genetics of dinucleotide (dC-dA)n . (dG-dT)n polymorphisms in world populations. *American Journal of Human Genetics.* **56**:461-474.

Deshpande C.H. *Academic American Encyclopedia.* Grolier Inc., Dunbury, Connecticut, **Vol. 14** . 1994.

DiRienzo A., Peterson A.C., Garza J.C., Valdes A.M., Slatkin M., and Freimer N.B. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA.* **91**: 3166-3170.

Dutta P.C. Biological anthropology of bronze age Harappans: New perpectives. (1983) *In*: "*The People of South Asia: The biological anthropology of India, Pakistan, and Nepal,*" Lukacs J.R., ed., Plenum Press, New York.

Edwards A., Civitello A., Hammond H.A., and Caskey C.T. (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics.* **49**: 746-756.

Edwards A., Hammond H.A., Jin L., Caskey C.T., and Chakraborty R. (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics.* **12**: 241-253.

Embree A.T. ed. *Encyclopedia of Asian History.* Charles Scribner's Sons, New York, **Vol. 2.** 1988.

Emeneau M.B. *Encyclopedia Americana* **Vol.9.** 1994.

Felsenstein J. (1993) *PHYLIP (Phylogeny inference package) 3.5c.* University of Washington.

Franda M., and Franda V.J. *Academic American Encyclopedia* GrolierInc., Dunbury, Connecticut, **Vol. 11.** 1994.

Garza J.C., Slatkin M., and Freimer N.B. (1995) Microsatellite allele frequencies in human and chimpanzees, with implications for constraints on allele size. Molecular Biology and *Evolution.* **12(4)**: 594-603.

Goldstein D.B., Linares A.R., Cavalli-Sforza L.L. and Feldman M.W. (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA.* **92**: 6723-6727.

Goren A., ed., *The Encyclopedia of the Peoples of the World.* (1993) Henry Holtz Co., New York.

Hammond H.A., Jin L., Zhong Y., Caskey C.T. and Chakraborty R. (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. *American Journal of Human Genetics.* **55**: 175-189.

Jorde L.B., Bamshad M.J., Watkins W.S., Zenger R., Fraley A.E., Krakowiak P.A., Carpenter K.D., Soodyall H., Jenkins T. and Rogers A.R. (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *American Journal of Human Genetics.* **57**: 523-538.

Kamboh M.I. (1984) Population genetic studies of PI, Tf, Gc and PGMI subtypes among various caste groups in North India. *Acta Anthropogenetica.* **8(3&4)**: 159-179.

Kirk R.L. (1976) The legend of Prince Vijaya - a study of Sinhalese origins. *American Journal of Physical Anthropology.* **20**: 91-100.

Kolman C.J., Sambuughin N. and Bermingham E. (1996) Mitochondrial DNA analysis of Mongolian populations and implications for the origin of new world founders. *Genetics.* **142**: 1321-1334.

Lahermo P., Sajantila A., Sistonen P., Lukka M., Aula P., Peltonen L. and Savontaus M-L. (1996) The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *American Journal of Human Genetics.* **58**: 1309-1322.

Levinson G. and Gutman G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution.* **4(3)**: 203-221.

Maciukenas M. (1991) *Treetool 1.0.* University of Illinois board of trustees.

Majumder P.P. and Mukherjee B.N. (1993) Genetic diversity and affinities among Indian populations: an overview, *in*: "*Human Population Genetics,*" Majumder P.P., ed., Plenum Press, New York.

Majumdar R.C. *Encyclopedia Americana. Deluxe Library Edition.* Grolier Inc., Dunbury, Connecticut. **Vol. 14.** 1994.

Minch E. (1996) *Microsat 1.4* . Stanford University.

Mountain J.L., Hebert J.M., Bhattacharyya S., Underhill P.A., Ottolenghi C., Gadgil M. and Cavalli-Sforza L.L. (1995) Demographic history of India and mtDNA-sequence diversity. *American Journal of Human Genetics.* **56**: 979-992.

Mukherjee B.N., Majumder P.P., Malhorta K.C., Das S.K., Kate S.L. and Chakraborty R. (1979) Genetic distance analysis among nine endogamous population groups of Maharashtra, India. *Journal of Human Evolution.* **8**: 567-570.

Murray B.W. (1996) The estimation of genetic distance and population substructure from microsatellite allele frequency data. McMaster University (not published). Available at: http://helix.biology.mcmaster.ca/brent/brent.html.

Nauta M.J. and Weissing F.J. (1996) Constraints on allele size at microsatellite loci: Implications for genetic differentiation. *Genetics.* **143**: 1021-1032.

Nei M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics.* **89(3)**: 583-590.

Nei M. and Roychoudhury A.K. (1993) Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution.* **10(5)**: 927-943.

Nei M. and Takezaki N. (1996) The root of the phylogenetic tree of human populations. *Molecular Biology and Evolution.* **13(1)**: 170-177.

Papiha S.S., Mukherjee B.N., Chahal S.M.S., Malhorta K.C. and Roberts D.F. (1982) Genetic heterogeneity and population structure in north-west India. *Annals of Human Biology.* **9(3)**: 235-251.

Passarino G., Semino O., Bernini L.F. and Santachiara-Benerecetti A.S. (1996) Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. *American Journal of Human Genetics.* **59**: 927-934.

Promega Technical Manual-TMD004 (1996) GenePrint™ STR Systems. Promega Co. Madison, Wisconsin.

Promega Technical Manual-TM023 (1996) Silver Sequence™ DNA Sequencing System. Promega Co. Madison, Wisconsin.

Queller D.C., Strassmann J.E. and Hughes C.R. (1993) Microsatellites and kinship. *Trends in Ecology and Evolution.* **8(8)**: 285-288.

Radhakrishnan R. *Academic Americana Encyclopedia* . **Vol. 6. 1994.**

Reynolds J., Weir B.S. and Cockerham C.C. (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics.* **105(3)**: 767-779.

Roewer L., Nagy M., Schmidt P., Epplen J.T. and Herzog-Schroder G. (1993) Microsatellie and HLA class II oligonucleotide typing in a population of Yanomami Indians, *in*: "*DNA Fingerprinting: State of Science*," Pena S.D.J., Chakraborty R., Epplen J.T. and Jeffreys A.J., ed., Birkhauser Verlag Basel, Switzerland.

Roff D.A. and Bentzen P. (1989) The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. *Molecular Biology and Evolution.* **6(5)**: 539-545.

Roychoudhury A.K. (1977) Gene Diversity in Indian Populations. *Human Genetics.* **40**: 99-106.

Roychoudhury A.K. (1983) Genetic relations between Indian populatoins and their neighbors, *in*: "*The People of South Asia: The Biological Anthropology of India, Pakistan and Nepal*," Lukacs J.R., ed., Plenum Press, New York.

Roychoudhury A.K. and Nei M. (1985) Genetic relationships between Indians and their neighboring populations. *Human Heredity.* **35**: 201-206.

Rubinsztein D.C., Amos W., Leggo J., Goodburn S., Jain S., Shi-Hua L., Margolis R.L., Ross C.A. and Ferguson-Smith M.A. (1996) Microsatellite evolution - evidence for directionality and variation in rate between species. *Nature Genetics.* **10**: 337-343.

Ruvolo M. (1996) A new approach to studying modern human origins: hypothesis testing with coalescence time distributions. *Molecular Phylogenetics and Evolution.* **5(1)**: 202-219.

Saha N., Tay J.S.H., Roy A.C., Das M.K., Das K., Roy M., Dey B., Banerjee S. and Mukherjee B.N. (1992) Genetic study of five populations of Bihar, India. *Human Biology.* **64(2)**: 175-186.

Saiki R.K., Gelfand D.H., Stoffel B., Scharf S.J., Higuchi R., Horn G.T., Mullis K.B. and Erlich H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science.* **239**: 487-491.

Saitou N. and Nei M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology of Evolution.* **4(4)**: 406-425.

Sanghvi L.D. (1976) Comparative genetic studies between some groups of Australian Aboriginals and certain tribal peoples of India, *in*: "*The Origin of the Australians*," Kirk R.L. and Thorne A.G., ed., Canberra: Australian Institute of Aboriginal Studies.

Schlotterer C. and Tautz D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Research.* **20(2)**: 211-215.

Shriver M.D., Jin L., Boerwinkle E., Deka R., Ferrell R.E. and Chakraborty R. (1995) A novel measure of genetic distance for highly polymorphic tamdem repeat loci. *Molecular Biology and Evolution.* **12(5)**: 914-920.

Shriver M.D., Jin L., Chakraborty R. and Boerwinkle E. (1993) VNTR allele frequency distributions under the Stepwise Mutation Model: A computer simulation approach. *Genetics.* **134**: 983-993.

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

Sokal R.R. and Rohlf F.J. *Biometry.* W.H. Freeman and Co., San Francisco, 1969.

Tammita-Delgoda S. *A Traveller's History of India.* Interlink Publishing Group Inc., New York, 1995.

Tautz D. and Renz M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research.* **12(10)**: 4127-4138.

Trabetti E., Galavotti R. and Pignatti P. (1993) Genetic variation in the Italian population at five tandem repeat loci amplified *in vitro*: use in paternity testing. *Molecular and Cellular Probes.* **7**: 81-87.

Walter H. (1971) Notes on the distributions of serum protein polymorphisms in India,*in*: "*Proceedings of the International Symposium on Human Genetics,*" Chakravartti M.R. ed., Andhra University Press, Waltair.

Walter H. (1986) Genetic differentiation processes among the populations of India. *International Journal of Anthropology.* **1**: 297.

Weber J.L. and Wong C. (1993) Mutation of human short tandem repeats. *Human Molecular Genetics.* **2(8)**: 1123-1128.

Weissenbach J., Gyapay G., Dib C., Vignal A., Morissette J., Millasseau P., Vaysseix G. and Lathrop M. (1992) A second-generation linkage map of the human genome. *Nature.* **359**: 794-801.

Zaykin D. and Pudovkin A. (1991) *CHIRXC, CHIHW.* Institute of Marine Biology. Vladivostok, Russia.