

ANCESTRY OF MODERN INDIAN POPULATIONS

ANCESTRY OF MODERN INDIAN POPULATIONS

By

AARON FRANCIS THOMSON, 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

© Copyright by Aaron Thomson, March 1999

MASTER OF SCIENCE (1999)
(Biology)

MCMASTER UNIVERSITY
Hamilton, Ontario

TITLE: Ancestry of Modern Indian Populations

AUTHOR: Aaron Francis Thomson, B.Sc.
(McMaster University)

SUPERVISOR: Professor R.S. Singh

NUMBER OF PAGES: ix, 135

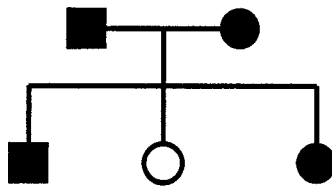
Abstract

An analysis of mitochondrial DNA (mtDNA) restriction fragment length polymorphisms (RFLPs) was done with the primary goal of clarifying the relationship of the Indian population to world populations. Phylogenetically informative RFLP sites were amplified, restricted and scored for 187 Indian-descended individuals. This sample collectively represented a geographically wide distribution, the three main religions present in the subcontinent, and three main caste groups. Thirteen haplotypes were found in the Indian population, and when combined with world population data obtained from the literature, 41 different haplotypes were found. India was found to be significantly different from all world populations under study. In agreement with previously reported results, the Indian population was found to be more similar to European populations than south-east Asian populations, with all Indian populations sharing the European-associated haplotypes 14 and 15 at high frequencies. However, high frequencies of haplotype 30 implied similarity with the Evenk population of Siberia, suggesting a possible north-central Asian origin for the Dravidian and/or Indo-Aryan migration into India.

Significant geographical differentiation within India was found, with north-western India having significantly higher frequencies of haplotypes 14 and 15 than Southern India, and lower frequencies of the Evenk-associated haplotype 30. The north-west was also significantly more diverse than other regions of India, most likely due to its location on the main routes of repeated migration into India. Significant differences

between religious groups were found to have a geographical basis, while caste groups were undifferentiated from each other and the main religious groups.

Dedicated to:



I wouldn't be who I am or where I am without you.

Acknowledgements

A two and a half year project can't be completed without owing thanks to many people. First among these are my supervisors / advisors, Dr. Rama Singh and Dr. Richard Morton. Without their support and project direction, and insightful questions and comments, this thesis would be less useful, interesting, and possibly less accurate.

Next on the list are Anouk Behara, Sujatha Thampi, and Harpreet Chana. Their efforts to gather samples (and in Harpreet's case analyzing the RFLPs of the ones that she gathered) are what made this project feasible.

Certain people in the Biology Department, most notably Sonya Grewal and Fariborz Yazdani, attempted to help me out in the initial phases of the project, when it seemed like I'd never get the PCR working. While the problems turned out to not be any of those that they had suggested, their time and advice was appreciated.

Rob Kulathinal, Pat Hayward and several of the secretaries have been around to help with any and all minutiae that are involved in the life of a graduate student. While no one incident stands head and shoulders above the rest, the general support with posters, TAing, lab equipment and advice for dealing with Life Science Building faculty and staff have made things easier for me than they otherwise would have been.

A nod must go to Rob Kulathinal (a slightly different aspect of him) and Lara Skwarek, who kept me in touch with what was going on around me and kept me from thinking that the universe was contained within the walls of the Life Science Building. In the same vein, merci to Craig Kelly and Usha Maharaj (CrUsha), Tammy Nadeau, and the gaming group for helping to fulfill my motto of "Work hard – play hard".

Table of Contents

Introduction	1
The Caste System	11
Muslim Influence in India	12
Sikh Origins	13
Christian Origins In India	13
Population Structure and Affinities: Blood and Protein Markers	14
Molecular Markers	18
This Study	22
Materials and Methods	23
Samples	23
DNA Extraction	23
RFLP Analysis	23
Data Analysis.	30
Results	40
Similarities Between Indian and World Populations	40
Statistical Tests	52
Populations Within India	56
Statistical Tests	61
Discussion	76
India and World Populations	76
Population Structuring Within India	84
Religious and Caste Differentiation Within India	86
Conclusions	89
Appendices	90
Appendix A – Protocols and Solutions	90
Appendix B – Indian Sample Restriction Data	101
Appendix C – Sorted Total Data	106
References	129

List of Tables

Table 1: Classical marker clines observed within India	17
Table 2: Geographical distribution of Indian individuals examined in this study	24
Table 3: Religious and caste composition of the Indian sample used in this study	25
Table 4: Linguistic composition of the Indian sample used in this study	25
Table 5: PCR primers and protocols	31
Table 6: Restriction enzymes used in this study, with associated data	33
Table 7: Origin of non-Indian RFLP data used in analyses.	34
Table 8: World Haplotypes	41
Table 9: Contingency table analysis of haplotype frequency distributions between India and nearest external populations	53
Table 10: Mitochondrial haplotype diversities of world populations	54
Table 11: Pairwise contingency table analysis of haplotype frequency distributions between European and African populations.	55
Table 12: Pairwise contingency table analysis of haplotype frequency distributions between Siberian and Korean populations, and within Siberia	57
Table 13: Haplotype ages as determined from world literature data	58
Table 14: Indian Haplotypes	59
Table 15: Indian Haplotype Frequencies	60
Table 16: Contingency table analysis of haplotype frequency distributions between areas within India – Significant Tests	70
Table 17: Contingency table analysis of haplotype frequency distributions between areas within India – Non-Significant Tests	71
Table 18: Contingency table analysis of haplotype frequency distributions between religions present within India	73

List of Figures

Figure 1: India and surrounding countries	3
Figure 2: Elevations within the Indian subcontinent	5
Figure 3: Distribution of samples and their religious affiliations	27
Figure 4: Distribution of Hindu samples with respect to caste	29
Figure 5: Subsections of the Indian subcontinent under study	37
Figure 6: Haplotype frequencies for global populations examined	43
Figure 7: Fitch-Margoliash distance trees of India and world populations	45
Figure 8: Minimum length majority-rules maximum parsimony consensus tree of world haplotypes with percentage resampling values	49
Figure 9: Maximum parsimony consensus tree with world haplotype distributions	51
Figure 10: Fitch-Margoliash distance trees of Indian subsections and world populations	63
Figure 11: Comparison of haplotype frequencies within religious groups and caste groups	65
Figure 12: Maximum parsimony consensus tree with Indian religious distributions .	67
Figure 13: Maximum parsimony consensus tree with Indian caste distributions . .	69

Introduction

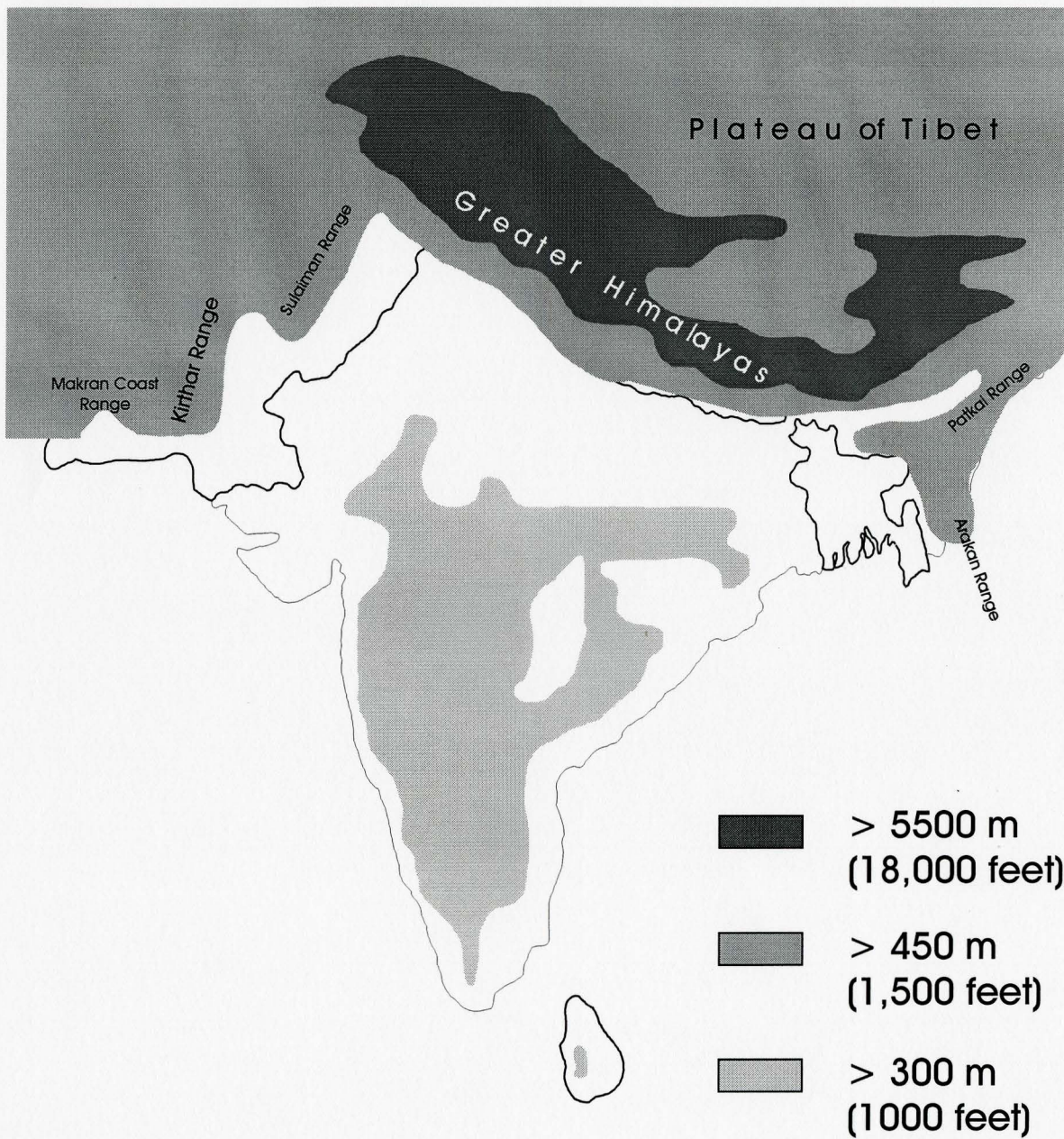
Curiosity about human origins has propelled scientific research for over a century. Over this time period, ethnic and geographical groups have been investigated and re-investigated using the most advanced scientific tools available, ranging from fingerprint, dental and craniometry analysis, to protein electrophoresis, to modern DNA-based molecular markers. Earlier studies in some cases have been shown by later studies to either be incorrect or wrongly interpreted (for example, the Torroni *et al* (1998) conclusions with respect to southern Europe as compared to those in Cavalli-Sforza *et al* (1994)). Therefore, periodic re-examination of conclusions is necessary in order to correct any mistakes. In this thesis, we present new data and conclusions on the origins and distribution of the populations of the Indian subcontinent, the home of one of the oldest extant cultures on the planet.

The Indian subcontinent is located at the south of the Asian continent, with the Himalayan mountains forming its northern border. The subcontinent is mainly composed of a triangular peninsula extending approximately 1,600 km into the Indian Ocean (See Figure 1). Mountain ranges caused by continental drift exist along many of India's borders with other countries. The Himalayan mountains directly to the north are the most formidable barrier between the subcontinent and the rest of Asia, but mountain ranges also exist in the north-east and the north-west (See Figure 2). However, mountain passes in the northwest and north-east lead to Afghanistan and Burma, respectively.

Figure 1: India and surrounding countries



Figure 2: Elevations within the Indian Subcontinent



The first evidence of hominid habitation in India occurs in Middle Pleistocene deposits – approximately 500,000 years Before Present (BP). Objects dated to this time period are usually primitive stone tools (Ghosh, 1989). However, one partial hominid skull dated between 200,000 to 700,000 years B.P. was found near Hathnora on the Narmada River floodplain, in the state of Madhya Pradesh (Lumley and Sonakia, 1985). Two groups of investigators have alternately identified the skull as an “evolved form” of *Homo erectus* (Lumley and Sonakia, 1985) and as an early form of *Homo sapiens* (Kennedy *et al*, 1991). However, comparisons to characterized European and African hominids of the same time-period could be misleading, as a 1995 hominid fossil find from the Central Narmada Valley implies that the Narmada hominid may have been pygmy-sized (Sankhyan, 1997). Due to the uncertain classification of the Narmada remains, it is unknown when modern *Homo sapiens* first entered the subcontinent, but the earliest definitely anatomically modern *Homo sapiens* remains that have been found date to the late Pleistocene (approximately 16,000 years old). These were found in a cave in Sabaragamuva Province, Sri Lanka, and are clearly modern *Homo sapiens* (Kennedy *et al*, 1987). The skeletons were slightly more robust cranially and postcranially than modern populations (Kennedy *et al*, 1987), which included prominent markers of muscle attachment and relative increases in bone size. Only fragments of older skeletons have been recovered (Kennedy and Deraniyagala, 1989). Unfortunately, these and other late Pleistocene and Mesolithic remains have been studied morphologically, but have not been studied using molecular markers to determine their relationship with extant populations (Ghosh, 1989).

In the historical period, the most significant migrations into the subcontinent have occurred in the northwest and north-east (Roychoudhury and Nei, 1985). As it is known that many different migrations into the subcontinent have occurred, archaeologists and anthropologists have attempted to reconstruct the histories of the many groups inhabiting the region. Efforts by anthropologists were initially based upon the religions, physical appearance, languages and marriage systems of the groups under study, but now also utilize protein and DNA markers.

The tribal groups of India (the scheduled tribes) compose about 7% of the total population of India and are divided into over 400 tribes of varying size (Cavalli-Sforza *et al*, 1994). Some of these tribal populations are thought to be the descendants of the earliest (surviving) groups in the subcontinent, while others are recognizable as more recent immigrants (Cavalli-Sforza *et al*, 1994). The tribal groups are usually those which are outside the typical social system (castes) of the majority of the population, but are not part of any otherwise recognizable group (e.g. Moslems or Christians). Genetic studies of Indian tribes have shown that they are usually most closely related to other tribes, followed by Indian non-tribal populations, and lastly world populations (Cavalli-Sforza *et al*, 1994). Protein and blood markers have shown that tribal populations are not actually related to the Australian Aboriginal or African populations which they physically somewhat resemble (Roychoudhury, 1984), but this only excludes two possible origins. Due to similarities in haplotypes of the hemoglobin β gene cluster containing the hemoglobin β^S (sickle cell) allele, it was proposed that some Indian tribal populations share a common origin, notwithstanding their present large geographical separation

(Lapie *et al*, 1989). The authors believed that this was consistent with the dispersion of a pre-existing population by more recently immigrating groups.

The first large group which is thought to have migrated into the subcontinent after the tribal groups is the Dravidians; possible descendants of this group are today mainly identified by the use of languages of the Dravidian family. Although the location of their original homeland is uncertain, this group is believed to have entered the subcontinent through the northwestern passes (Balakrishnan, 1978). Today, the Dravidian languages are spoken mainly in the four southern provinces of India, with pockets of Dravidian-speakers elsewhere (Cutler, 1988).

The Dravidians are thought to have been one of the culturally dominant groups in India at the time of the Bronze Age Harappan Culture of the Indus Valley (ca. 2500 BC - 1700 BC), as many place names outside of southern India and “borrowed” words in Indo-Aryan languages show signs of a Dravidian origin (Southworth, 1990; Southworth 1996). Unfortunately, the written language of the Harappan Culture has not been translated (Shaffer, 1998), limiting the amount we can learn. The civilization extended over 1600 km of the Indus Valley (which is located mainly within modern-day Pakistan near the border with India), making it the largest culture in the world at the time. Archaeologists have postulated many theories to explain the end of the Harappan Culture, but a consensus has emerged that (relatively gradual) ecological changes of either human or natural origin disrupted Harappan food production and trade (Shaffer, 1998). However, it is clear that while the largest Harappan cities may have been abandoned, the underlying society continued for some time and contributed elements as diverse as measurement

systems and religious influences (Agrawal, 1996) to cultures that have continued to the present day. Aryans migrating into India with higher (Iron Age) technology may have come into contact with the remnants of the civilization (Cavalli-Sforza *et al*, 1994; Schmidt, 1995), but the Aryans could not have been completely responsible for the fall of the civilization, as it appears to have been in serious decline at when the first indications of the Aryans occur in the archaeological record.

The largest group to enter India (by approximately 1000 BC) after the Dravidians is termed the Aryans, and is generally agreed to have contributed significantly to the modern culture and genetic structure of India (Cavalli-Sforza *et al*, 1994). Descendants of these Aryans (Indo-Aryans) speak languages which are related to European languages (of the Indo-European language family) (Cavalli-Sforza *et al*, 1994). An earlier consensus that the Aryans originated to the west or northwest of northern India has been slightly weakened by scholars who have noted that there is little archaeological evidence for an external origin, as opposed to a sudden expansion of an indigenous group (Shaffer, 1984). However, linguists generally believe that the Indo-European language family originated outside India (Southworth, 1990), as India is on the periphery of the historical range of the Indo-European language family.

While it is not clear when the Aryans began migrating into the Indian subcontinent, by 1000 BC they controlled much of the territory that had been Harappan. Increased usage of iron tools after 800 BC (Schmidt, 1995) and horse-mounted military units (Cavalli-Sforza *et al*, 1994) allowed the Indo-Aryans to eventually form the dominant culture throughout most of the Indian subcontinent. Evidence of this includes

the fact that elements of Indo-Aryan culture such as the caste system and religious traditions are now found throughout India, and that Indo-European languages are now spoken by the large portions of the population in all parts of India (Cavalli-Sforza *et al*, 1994).

Subsequent invasions and occupations of parts of India by various groups are firmly dated in the historical record. These groups include: Greeks (550-50 BC), Sakas (Scythians) (ca. 80 BC) (Schmidt, 1995), Kushans (ca. 100 AD), Huns (AD 200-500) (Papiha, 1996), Arabs (1000-1206), Mongols (1526-1658), (Schmidt, 1995; Papiha, 1996) and lastly the Europeans (1510-1947). These groups are thought to have had less effect on the culture and gene pool of the continent than previous invasions, as most did not attempt to settle India, but instead simply conquered and replaced the existing dynasties. Most later invaders also did not conquer the entire subcontinent (Schmidt, 1995). Some groups (such as some Huns) did not completely assimilate into the local populations but became new tribal groups (India, 1998). Unfortunately, very little is known about migration and gene flow into the Indian subcontinent through the northeast; however, the presence of populations speaking Austroasiatic languages (the tribal Mundas, in central & northern India) and Sino-Tibetan languages (in the north-east) (Cavalli-Sforza *et al*, 1994) suggests that such migrations have taken place. Protein and blood marker studies for some groups in north-eastern India also show frequency similarities to populations outside of India (Papiha *et al*, 1996) which decrease with geographical distance from the border.

The Caste System

The caste system is believed to have been introduced into India by the Indo-Aryans, as it is mentioned (and justified as both god-given and due to innate traits) in the Vedas (traditional historical-religious texts of the Indo-Aryans) (Muir, 1868). As specified in the Vedas, there were originally four groups or *varnas*, in order of status: the Brahmins (priests and teachers), Kshatriyas (lords and elite warriors), Vaishyas (traders and farmers) and Sudras (servants or laborers). It is thought that in northern India the Dravidians were either incorporated into the ranking as Sudras or were excluded from the caste system entirely, eventually becoming the Harijan (untouchable) class (Caste, 1998). The caste system (with various sub-castes) was codified between 200 BC and 100 AD in the socio-religious text *Manu Samriti* (Law of Man) (Tambiah, 1973). From this point onward (and likely for some time previously) all castes were (by religious law) endogamous (i.e. only marriages within a caste were permitted) (Bamshad *et al*, 1996). For the upper castes, familial associations as laid down in the *Manu Samriti* were to be followed to determine permitted (*gotra*) marriages (R. Singh, per. comm.). However, enough exceptions to the caste rules occurred between socially adjacent castes that rules for sub-castes were included in the *Manu Samriti*, each of which had their own level of privilege and rules of allowable marriages (Tambiah, 1973). In general, if a man married a lower-caste woman, the children might retain his status, but would more likely have an intermediate status (within a sub-caste). If the father was of a lower caste than the mother, the children would possibly be assigned to his caste, and but might be excluded from the community as Untouchables (Tambiah, 1973). Therefore, there was a large

incentive for women not to marry lower caste husbands, beyond the usual rules for within-caste marriage. The over-all effect of this system was to both stabilize the Brahmins and Kshatriyas as the ruling classes (due to the Brahmin role as interpreters of the Vedas, and the second-highest status for the warriors given in the Vedas) and to reduce gene flow between castes.

Muslim Influence in India

After the establishment of small Islamic states in present-day Pakistan in 712-713, and a series of successful raids in 1000-1025 by Mahmud of Ghazni, the first Islamic empire in northern India was established in 1175-1206 by Muhammad Ghuri (Schmidt, 1995). This was the nucleus of the Delhi Sultanate, which by 1351 had gained control of almost the entire subcontinent. Internal factors undermined the central authority's control, and the southern border receded to the north, leaving some Muslim kingdoms in the south. After 1398, the Delhi Sultanate shrank to the Delhi area and after a brief resurgence was conquered by Turks, who founded the Mughal Empire in 1526. By 1707, almost the entire subcontinent (once again excluding the southernmost areas) was again under nominal central control (Schmidt, 1995). During the various periods of Islamic control of parts of the Indian subcontinent, Muslims formed the ruling class. During these time periods, many people in areas under the control of the Delhi Sultanate and Mughal Empire converted to Islam, which would have the effect of diluting the contribution by outsiders to the Muslim population in India (Papiha, 1996).

Sikh Origins

The Sikh religion was founded by the mystic Nanak (1469-1539) as a fusion of Islam and Hinduism. During the religion's initial period of growth (by conversion from established religions), a military tradition was established to oppose persecution from the religious majority. This led to the creation of a Sikh state in the Punjab by about 1800 which was eventually incorporated into British India (Sikhs, 1998). By 1961, 7.8 million Sikhs were living in India, 86 percent of whom lived in Punjab. Tradition in Punjab for some time mandated the conversion of first-born sons from Hinduism to Sikhism (R. Singh, per. comm); therefore genetic differences between the Sikh and local Hindu population would not be expected.

Christian Origins in India

The Christian church is known to have existed in Kerala since at least 525 A.D., as it was reported there and in Sri Lanka by the Greek writer Cosmas Indicopleustes (Neill, 1970). This early church was reported as being composed of Persians (presumably merchants) with Persian clergy, presumably residing there due to the commonly utilized sea trade route between Egypt and Southern India (Neill, 1970). The earliest Christian churches in Kerala have traditionally believed that they were founded by conversion of high caste Hindus by Thomas the Apostle in the first century A.D., but there is little supporting evidence for this.

The advent of Islam in the Middle East severed almost all contact between India and Europe until the conquest of Goa by the Portuguese in 1510 A.D. Mixed marriages between male traders and local women began almost immediately, but a larger increase in

Christians occurred with the mass conversion of an entire local caste on the eastern coast of Tamil Nadu in approximately 1536. The original Christian churches in Kerala underwent a European-caused schism, with approximately 2/3 joining the Roman Catholic Church, and 1/3 remaining independent. In later centuries, an overall lack of translation and accommodation of the Christian faith to the Indian culture had slowed conversions such that by 1800 only about half a million Christians lived in India, mainly in Goa and Kerala. However, after this time missionary efforts (and the ordination of Indians) resulted in a steady trickle of converts, such that by 1961 there were approximately 11 million Christians in India (over half of which lived in Kerala, Tamil Nadu and Andhra Pradesh) (Gupta, 1961).

Population Structure and Affinities: Blood and Protein Markers

The repeated invasions of the Indian subcontinent (especially those of the Dravidians and Aryans) have had visible effects on the present genetic structure of India, as shown by studies with classical blood groups, serum proteins and red cell enzymes. These effects include: 1. Similarities between north-eastern populations (Austro-Asiatic and Tibeto-Chinese speakers), some tribal populations and south-east Asia; 2. Similarities between north-western populations, Middle Eastern and European populations, and also 3. Gene frequency clines from the north-western states to the southern and eastern states (Papiha, 1996).

Physical similarities between populations in north-eastern India and those in south-east Asia had been noted by anthropologists before the advent of classical markers. Allele frequency comparisons of blood group markers *MNS* allele *M* (*MNS***M*),

*MNS*MS*, Rhesus allele *D* (*Rh*D*), *Rh*CDe*, *Rh*Cde*, *PI*1*, and the 6-phosphogluconate dehydrogenase allele *C* (*PGD*C*) support this relationship (Cavalli-Sforza *et al*, 1994). It has also been noted that Duffy (*FY*) allele *A* frequency is lower than 50% in Indian Austro-Asiatic speaking populations, which is midway between the frequencies found in south-eastern Asian populations and Indian populations (Papiha, 1996). Some tribal populations have also been shown to have similarities to south-east and central Asians with respect to some blood group markers, in particular *MNS*N*, *Rh*D*, *Rh*Cde* and *Rh*CDe*. Acid phosphatase (*ACPI*) allele *C* is also missing from the tribal populations that have been sampled thus far, but this perhaps simply shows that there has been little European contribution to these gene pools as this allele is only relatively common (4-8%) in Europe (Cavalli-Sforza *et al*, 1994).

The north-western population of the subcontinent shows definite similarity with populations farther west or north-west with respect to the frequencies of the blood group markers *FY*A*, *FY*B*, *Rh*D*, *TF*C*, vitamin-D binding protein (*GC*1*) and *ACPI*A* (Papiha, 1996; Cavalli-Sforza *et al*, 1994). Similarities are also present with other markers and alleles such as *ABO*O*, *ACPI*B*, *ACPI*C*, *ADA*1*, *GLO1*1*, *GC*1*, *GC*1F*, *HCAA*9*, *HCAA*10*, *HLAB*18*, *MNS*S*, *Rh*cde* (Cavalli-Sforza *et al*, 1994). However, similarities with these loci and alleles are less certain due to factors such as insufficient sampling density, similar frequencies being found in populations other than in the west or north-west, continent-wide low levels of variation in allele frequency, or the pattern observed with respect to India being part of a larger continent-wide pattern (e.g. a frequency cline from Europe to China).

A large number of classical markers show gene frequency clines (possibly implying gene flow) within India itself (Papiha, 1996). Table 1 shows a summary of cline directions within India and whether they are a part of a larger pattern. In summary, it appears that classical markers give a picture of allele frequency clines running either east-west, or northwest-southeast. Excluding small areas of differing gene frequency in the north-east, no clines appear to run from the northeast to southwest, as would be expected if a major proportion of the Indian gene pool had entered at the northeast of the subcontinent. Most within-India trends (with the exceptions of *HP*1*, *ABO*O*, *ABO*B* and *Rh*Cde*) appear to be part of overall Eurasian gene frequency trends, with (for some alleles) Middle East-characteristic frequencies extending into India, farther east than they reach elsewhere (Cavalli-Sforza *et al*, 1994).

Phylogenetic analysis of classical markers on a world-wide scale usually places Indian populations as a group relatively closely to European populations (Cavalli-Sforza *et al*, 1994), with some notable exceptions appearing in a minority of bootstrapping repetitions relating Dravidian populations to East Asians (Cavalli-Sforza *et al*, 1988). Within-India phylogenetic studies using classical markers have had differing results dependent upon the sampling location. Populations in the state of Assam have been shown to cluster based on geographical proximity, implying that the sociocultural hierarchy (of castes and other groups) in this area has not had a large effect on these groups (Majumdar and Mukherjee, 1993). However, for populations from northwest India genetic affinity is somewhat correlated with close sociocultural grouping (Papiha, 1996). An analysis of populations from Uttar Pradesh, Andhra Pradesh and Gujarat also

Table 1: Classical marker clines observed within India

Allele or Haplotype	Cline direction	Part of a larger Eurasian pattern?
<i>ABO*B</i>	decreasing N to S	no
<i>ABO*O</i>	increasing N to S	no
<i>ABO*A</i>	increasing NW to NE	no
<i>Rh*Cde</i>	increasing NW to SE	no
<i>Rh*CDe</i>	increasing W to E	yes
<i>Rh*cde</i>	decreasing W to E	yes
<i>MNS*MS</i>	decreasing NW to NE	yes
<i>FY*A</i> ¹	increasing NW to S	yes
Haptoglobin allele <i>1</i> (<i>HP*1</i>)	decreasing NW to SE	yes
Adenosine deaminase allele <i>1</i> (<i>ADA*1</i>)	increasing NW to E	yes
Acid phosphatase <i>1</i> (<i>ACP1*A</i>)	decreasing NW to SE	yes
<i>GC*IF</i>	increasing W to E	yes
<i>GC*1</i>	increasing N to S	yes
<i>AKI*1</i>	increasing NW to NE	yes

¹There are small areas of low *FY*A* frequency in the northeast, corresponding to Indo-Aryan and Austro-Asiatic speakers (Papiha, 1996).

Sources: Cavalli-Sforza *et al*, 1994; Papiha, 1996.

showed social clustering, with the (geographically widely distributed) Muslim groups being closest together, followed by Hindu caste groups (which were sampled from Uttar Pradesh), then tribal groups from Andhra Pradesh. However, in a 6 state Hindu / Muslim comparison, Muslims and Hindus mainly clustered geographically, implying that temporary Islamic social dominance did not alter the genepool of the subcontinent substantially (Papiha, 1996). The “broad generalizations” that have emerged with respect to the caste system appear to be that a) within some regions genetic distances are substantial between caste groups, b) caste groups belonging to different varnas (main caste groups) often exhibit considerable genetic distances (especially between the castes at the top and bottom of the socio-cultural hierarchy), and c) geographically clustered castes are more similar when compared to groups in other areas irrespective of social hierarchy (Bamshad *et al*, 1996). Given the varying results and continually accumulating data, periodic reviews of the protein and blood marker data would be welcome.

Molecular Markers

In general, molecular markers such as mitochondrial DNA, Y chromosome markers and autosomal markers have become widely used for purposes of phylogenetic analysis. Any molecular marker such as RFLPs, microsatellites or sequences from non-coding DNA is relatively advantageous in that it is almost certainly more neutral than classical markers with respect to selection. Mitochondrial DNA was the first of these to be widely used, mainly due to its high copy number and consequent ease of isolation (Awise *et al*, 1987) and also the fact that it had been completely sequenced (Anderson *et*

al 1981). Other useful mtDNA traits include a) a simple genetic structure with few genes, small intergenic regions and the “D-loop” transcription control region; b) clear transmission (mother to children), without the recombination that nuclear markers may undergo; and c) rapid evolution allowing sequence differences to arise frequently (Avisé *et al*, 1987). Disadvantages of mtDNA with respect to human phylogenetic studies, are rare occurrences of more than one genotype within a single person, and parallel or convergent mutations of some “hypervariable” restriction sites / nucleotides within the population (Avisé *et al*, 1987). Exploitation of autosomal and Y chromosome molecular markers did not generally begin until after the introduction of the polymerase chain reaction.

Studies of the Indian population using molecular markers such as mitochondrial DNA (mtDNA) and the Y-chromosome have been done less frequently than those using classical markers. However, mtDNA comparisons with other world populations have shown a clear relationship between Indians and Caucasians, with some Asian admixture (Mountain *et al*, 1995; Barnabas *et al*, 1996). Also, mtDNA studies have shown (using high-resolution restriction analysis) lower frequencies of Caucasian markers in the south than in the north, as seen by 10,394 *DdeI*, 10,397 *AluI* and 7,025 *AluI* (Passarino *et al*, 1996). Alternatively, low-resolution mtDNA restriction analysis suggested the continuing presence of an older Caucasian migration in southern India which has been infused with a newer Caucasian migration in the northern parts of the subcontinent (Barnabas *et al*, 1996). Higher observed levels of nucleotide diversity in north-west and north-central India than central and southern India was consistent with this (Barnabas *et*

al, 1996), although population size can affect diversity. Lastly, a world-wide study of Y chromosome variation including the Kota (tribal), Madras and Sri Lankan populations has suggested a high affinity with respect to Y chromosome haplotype frequencies between South Indian and North Asian populations such as West Siberians and Kets, or possibly Central Asians such as the Altai and Mongolians (Hammer *et al*, 1998).

Studies of caste using molecular markers have had inconsistent results, even though only southern populations have been sampled. A study of mtDNA sequence (hypervariable segments (HVS) I and II) from Brahmin, Harijan and tribal groups (Mountain *et al*, 1995) showed some clustering according to caste, but no clear separations into distinct groups. A faster growth rate for the Brahmin population was inferred from differences in the pairwise nucleotide difference distributions, which corresponded to simulation studies of expanding populations. The neighbor-joining tree generated from the Mountain *et al* (1995) data and control data supported the interpretation that the common ancestor of the Indian lineages sampled predated the divergence of Eurasian populations, and that little gene flow between Indian and Eurasian populations occurred after their separation. It has more recently been suggested (Govindaraju, 1995) that sufficient gene flow has occurred since the founding of the caste system to obscure any original or consequent patterns of caste mtDNA variation.

A large study of mtDNA and Y-chromosomal variation in south Indian populations (Bamshad *et al*, 1998b) showed unimodal mismatch distributions as in Mountain *et al* (1995) but for each caste instead of just the Brahmin, suggesting large expansions (but before the last 2,000 years). The south Indian population in general was

found to have more affinity to Asian populations than Caucasians (Bamshad *et al*, 1998b), while some individuals in the upper castes were found to have more similarities with Caucasians (Gibbons, 1998). A later paper (Bamshad *et al*, 1998a) found a positive correlation between caste rank and mtDNA diversity (such that mtDNA diversity was highest in the upper castes), and that mtDNA genetic distance was highest between the highest and lowest castes. The diversity result is somewhat correlated with present caste size, as Brahmins are presently among the largest castes in some areas (Dobzhansky, 1962). Y-chromosomal genetic distance was not correlated with caste, leading the authors to conclude that the distance and diversity results could be best explained by female-specific gene flow between castes (Bamshad *et al*, 1998a).

Another study of 36 Hindu males of each of the four castes showed 25 mtDNA (HVS II) haplotypes, only three of which were shared between castes (Bamshad *et al*, 1996). Reflecting this, the reported G_{ST} value for the castes was 0.17 ($p < 0.002$) – meaning that 17% of the variance of the mtDNA region sampled occurred between caste groups. This G_{ST} value was not in agreement with previously reported values derived from blood and protein markers, but higher G_{ST} values are often noted for mitochondria due to differences in population size. This study also noted that the Indian population clustered (using a neighbor-joining algorithm on genetic distance matrices) more often with the African sample than with the Asian or European sample. This led the authors to suggest that admixture with African populations had occurred. However, subsequent analysis of Y chromosome variation in this population failed to find evidence for this

scenario (Spurdle *et al*, 1997) due to a lack of African-characteristic Y-chromosomal haplotypes and alleles (frequency 0.61-0.95 in Africans) in the Indian population.

This Study

To this point, no published study based on molecular markers has compared a geographically diverse sample from India to world populations. Therefore, the main objective of this study was to use molecular markers to attempt to determine the contributions of various world populations (especially Europeans and East Asians) to the genetic makeup of the entire Indian subcontinent, and geographical subsections of the subcontinent. Other pertinent questions, such as possible caste divisions, religious divisions and linguistic divisions were also examined. To this end, mtDNA restriction sites known to be phylogenetically useful as population-specific markers were scored and studied as haplotypes.

Materials and Methods

Samples

Blood samples were obtained from volunteers residing in Southern Ontario who were immigrants from the Indian subcontinent or first generation Canadians of this descent. Volunteers filled out informed consent forms, including information on name, gender, place of birth, parents' places of birth, and parents' native language. (See Table 2, Table 3, Table 4, Figure 3 and Figure 4.) Information on religion was optional, and caste was assigned by surname by knowledgeable individuals. Blood was drawn into 5 or 7.5 ml Vacutainers (containing 0.05 ml of 15% EDTA solution) by a qualified technician. Blood was stored in 1 ml aliquots at -70°C .

DNA Extraction

DNA was extracted using a standard cell-lysing and phenol extraction protocol. Red cells were lysed with a mixture of NH_4Cl and NH_4HCO_3 (repeated as necessary). Subsequently, white blood cells were lysed and incubated with a mixture of SSTE (0.5% SDS in Sodium-Tris-EDTA solution (STE)) and proteinase K. Standard phenol extraction and ethanol-salt DNA precipitation techniques were used (Sambrook et al, 1989). The DNA pellet was resuspended in 50 μl of TE (Tris-EDTA solution) and concentrations were quantified using a fluorimeter and intercalating dye.

RFLP Analysis

Anticipated restriction sites (based on Anderson *et al*, 1981) for each primer pair were planned in conjunction with picking primers, so that restriction fragments would be

Table 2: Geographical ancestry of individuals examined in this study.

Region and State (or Country)	# of samples
Southern	
Sri Lanka	1
Tamil Nadu	19
Kerala	13
Karnataka	4
Goa	1
Andhra Pradesh	2
	<u>40</u>
Northern	
North-West	
Pakistan	10
Jammu and Kashmir	1
Himachal Pradesh	1
Punjab	29
Haryana / New Delhi	6
Rajasthan	1
Gujarat	6
	<u>54</u>
North-Central	
Uttar Pradesh	43
Madhya Pradesh	1
Nepal	1
Bihar	3
	<u>48</u>
North-East	
West Bengal	20
Bangladesh	11
Assam	1
Meghalaya	1
	<u>33</u>
Maharashtra	7
Location Unknown	5
Total Samples	187

Table 3: Self-assigned religious composition and presumed caste composition of the samples used in this study.

Religion / Caste	# of Samples
Hindu	
Brahmin	36
Kshatriya	20
Vaisya	26
Unknown	<u>34</u>
	116
Sikh	32
Muslim	16
Christian	18
Jain	<u>3</u>
Total Known	185
Unknown	2

Table 4: Linguistic ancestry of the samples used in this study.

Language	# of samples
Dravidian	38
Indo-European	<u>147</u>
Total Known	185
Unknown	2

Figure 3: Distribution of samples and their religious affiliations by ancestry.
(As in Table 2.)

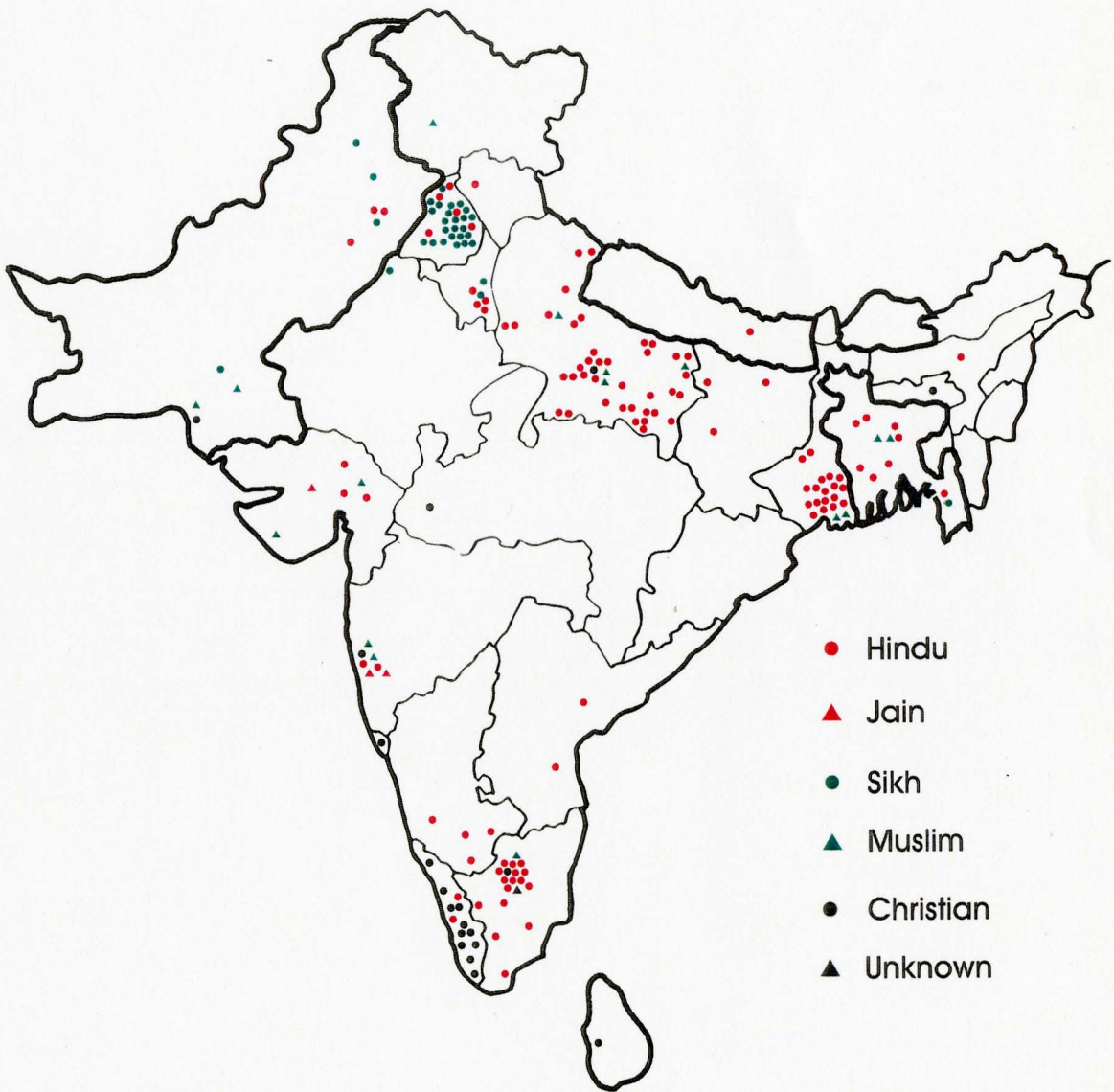
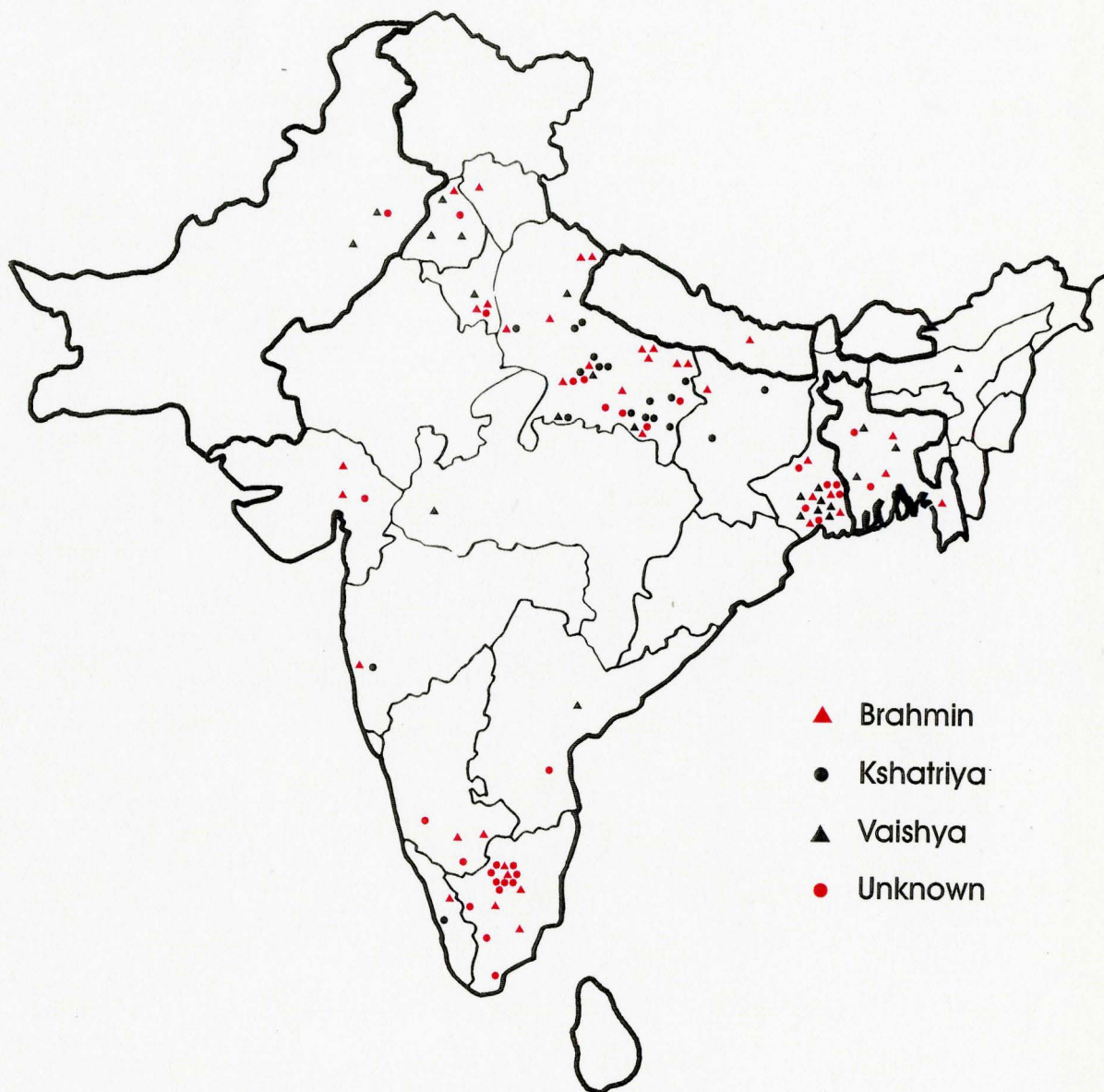


Figure 4: Distribution of Hindu samples with respect to caste.



clearly visible and (in the case of duplex reactions) would not conceal fragments of interest due to similar sizes. Length of the DNA segment to be amplified was kept under 400 base pairs to maximize yield. OLIGO version 3.4 (Rychlik and Rhoads, 1989) was used in all cases. (See Table 5 and Table 6.)

PCR was carried out as described in Appendix A. PCR products were run on 1-2% agarose minigels which were then stained with ethidium bromide and visualized under ultraviolet light to ensure that amplification had taken place. 5-15 μ l (dependent on band intensity seen on the initial gel) of PCR product was then restricted for 3-4 hours at 37°C using 2-4 units of commercially available restriction enzymes and buffers. The resulting fragments were run on 3-4% agarose minigels and visualized as above in order to score the presence or absence of restriction sites.

Data Analysis

Data resulting from this study and published data which included corresponding sites (see Table 7) were sorted into haplotypes. The haplotype frequencies for each population were quantified, and with the haplotypes (expressed as site presences/ absences) were used in the PHYLIP software package (Felsenstein, 1995) to accomplish phylogenetic analysis. India was separated into regions for some analyses (see Figure 5). Samples from the state of Maharashtra were not included in analyses that involved separating northern India into areas, as it was not clear to which area that these samples belonged. The state of Maharashtra was not grouped with South India due to the distinct linguistic nature (majority Dravidian-speaking) of the four southernmost states (see

Table 5: PCR Primers and Protocols

	Primers and Synthesis Number ¹	Mp (°C) ²
Set A 9911- 11873	9911-9932 AB9524 5',CGA AGC CGC CGC CTG ATA CTG G ₃ '	66.2
	11873-11851 AB9525 5',GG GGG GTA AGG CGA GGT TAG CG ₃ '	65.5
Set B 10349- 10558	10349-10370 AB10552 5',CCT AGC CCT AAG TCT GGC CTA ₃ '	52.2
	10558-10537 AB10553 5',GGG AGG ATA TGA GGT GTG AGC ₃ '	52.5
Set C 6952- 7225	6952-6972 AB10779 5',TAG GTG GCC TGA CTG GCA TTG ₃ '	60.1
	7225-7205 AB10778 5',GAG TAA CGT CGG GGC ATT CCG ₃ '	57.3
Set D 8150- 8366	8150-8167 AB11345 5',CCG GGG CTA TAC TAC GG ₃ '	47.4
	8366-8344 AB11346 5',TTT CAC TGT AAA GAG GTG G ₃ '	48.9
Set E 13224- 13366	13224-13245 AB11347 5',TGA CAT CAA AAA AAT CGT AGC ₃ '	47.5
	13366-13345 AB11348 5',GGA GCA CAT AAA TAG TAT GGC ₃ '	46.2
Set F 515- 727	515-535 AB12108 5',ACA CAC ACA CCG CTG CTA ACC ₃ '	54.9
	727-707 AB12109 5',AGG GTG AAC TCA CTG GAA CGG ₃ '	55.3

Set G 5130- 5324	5130-5150 AB12110 5'TTA AAC TCC AGC ACC ACG ACC ₃ '	54.1
	5324-5304 AB12111 5'GAT GGT GGC TAT GAT GGT GGG ₃ '	56.3
Set H 15977- 16170	15977-15997 AB12643 5'CCA CCA TTA GCA CCC AAA GCT ₃ '	55.6
	16190-16170 AB12644 5'GAG GGG GTT TTG ATG TGG ATT ₃ '	54.2
Set I 16399- 70	16399-16419 AB12645 5'ACC ACC ATC CTC CGT GAA ATC ₃ '	55.0
	70-50 AB12646 5'CCC CCA GAC GAA AAT ACC AAA ₃ '	55.4

¹ Numbering as in Anderson *et al*, 1980. All primers were purchased from the Central Facility of the Institute for Molecular Biology and Biotechnology, McMaster University.

² As predicted by OLIGO 3.4

Table 6: Restriction enzymes used in this study, with associated data.

Enzyme	Restriction Sequence	Restriction sites Examined	Located in gene:	Site useful for:	Percentage in this area	Percentage elsewhere
<i>AluI</i>	AG/CT	5,176	NADH dehydrogenase 1	Asia	72 - 100	99 - 100
		7,025	Cytochrome c Oxidase I	Europe	33 - 79	98 - 100
		10,397	NADH dehydrogenase 3	Asia	33 - 92	0 - 1
		13,262	NADH dehydrogenase 5	Asia	16 - 100	0
<i>DdeI</i>	C/TNAG	10,394	NADH dehydrogenase 3	-		
<i>HaeIII</i>	GG/CC	663	12S rRNA	Asia	0 - 11	0
		16,517	D-loop	-		
<i>HinfI</i>	G/ANTC	16,065	D-loop	Europe	90 - 91	99 - 100
- (9 bp deletion)	-	8272-8281	intergenic COII/tRNA ^{Lys}	Asia, Africa ¹	2-22,0-27	0 - 1

Data from Ballinger *et al*, 1992; Chen *et al*, 1995; Kogelnik *et al*, 1998; Kolman *et al*, 1996; Melton *et al*, 1995; Merriwether *et al*, 1996; Soodyall *et al*, 1996; Torroni *et al*, 1993a; 1993b; 1994a; 1994b; 1996; 1997; 1998

¹Independent origins of 9bp deletion can be shown in conjunction with other mtDNA data.

Table 7: Origin of non-Indian RFLP data used in analyses.

Origin	Subpopulations	# individuals	Study
Native American	Dogrib	30	Torrioni <i>et al</i> , 1993.
	Navajo	48	
	Ojibwa	28	
	Pima	30	
	Maya	27	
	Boruca	14	
	Kuna	16	
	Guaymi	16	
	Bribri-Cabecar	24	
	Yanomama	24	
	Piaroa	10	
	Makiritare	10	
	Macushi	10	
	Wapishana	12	
	Ticuna	28	
	Kraho	14	
	Marubo	10	
Mataco	28		
	<u>379</u>		
Siberian	Nivkhs	57	Torrioni <i>et al</i> , 1993.
	Evenks	51	
	Udegeys	<u>45</u>	
		153	
Tibetan	no subgroups	54	Torrioni <i>et al</i> , 1994.
Southeast Asians	Malaysian Chinese	14	Ballinger <i>et al</i> , 1992.
	Vietnamese	28	
	Malay Aborigines	32	
	Malays	14	
	Taiwanese Han	20	
	Koreans	13	
	Sabah Aborigines	<u>32</u>	
		153	

Africans	Western Pygmies	17	Chen <i>et al</i> , 1995.
	Eastern Pygmies	22	
	Niokolo Mendenkalu	60	
	Wolof	20	
	Pular	8	
	Other	<u>13</u>	
		140	
Italians	Italians	51	Torrioni <i>et al</i> , 1997.
	Sardinians	<u>48</u>	
		99	
Europeans	Finnish	49	Torrioni <i>et al</i> , 1996.
	Swedish	<u>37</u>	
		86	
North Americans (Caucasians)	French Canadian	28	Torrioni <i>et al</i> , 1994.
	Americans	<u>147</u>	
		175	

Figure 5: Subsections of the Indian subcontinent under study.

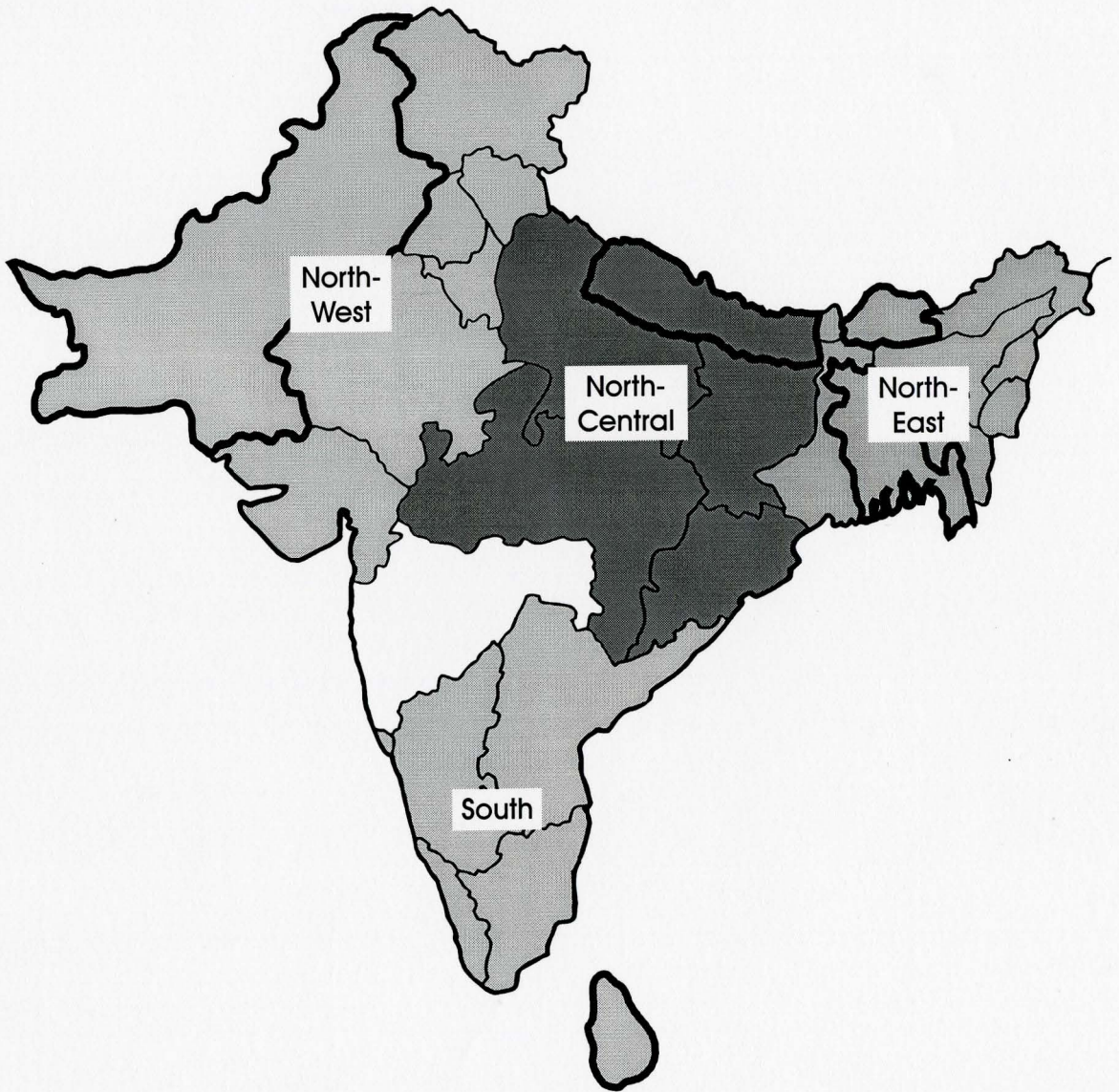


Figure 5). Haplotype frequencies were used to produce genetic distance tables using GENDIST. The Cavalli-Sforza chord distance option was used due to the fact that some populations had no alleles in common, a situation that some other distance algorithms cannot handle. These tables generated by GENDIST were used as input for tree-generating programs FITCH. The Global rearrangement option was turned on, so that subtrees would be removed from the generated tree and put back on in all possible ways (resulting in more trees being examined).

Dollo parsimony analysis of the haplotypes was accomplished using DOLLOP. A data file containing many random order repetitions (602) of the haplotype data was used to generate a file of minimum length trees (602 * 100 minimum length trees stored from each repetition = 60200 trees found) using the Global Rearrangement option and Jumble option (10 jumbles for each input tree). These trees were then used as input for CONSENSE to generate a majority rule, strict consensus tree.

This consensus tree was used as the basis for allocation of haplotypes into categories for statistical testing. Statistical tests for homogeneity of populations was done using the G-test and the chi-squared test. If the table being tested was in a 2 X 2 format, Fischer's Exact Test was also done; otherwise the table was tested using Monte Carlo simulation with 10000 runs in the CHIRXC program (Zaykin and Pudovkin, 1993). Significance was checked with the Sequential Bonferroni tests to avoid random significance at the 0.05 level (due to the number of tests being done).

The age of the two most common Indian haplotypes pairs (the haplotypes composing the pairs are distinguished only by the hypervariable 16,517 *HaeIII* site) were

estimated using literature data. The age of the predominant African haplotype pair was estimated, as was that of the most widespread haplotypes in East Asia. Ten samples of each haplotype were selected from each world population in which these haplotypes were present. Using the whole-mitochondria restriction data for these samples, pairwise mean divergence was estimated as described in Nei and Tajima (1983). Some samples from Central America were excluded from this analysis as they were not present in any other North American or East Asian populations, and were therefore suspected of being a possible example of parallel evolution in the sites defining the haplotypes. Similarly, if a sample was noticeably more divergent than others in the analysis, it was excluded from a second analysis in order to determine how much of the total divergence was due to that sample. Pairwise mean divergences were averaged to obtain the average mean divergence estimate (π). A time estimate was obtained by dividing π by a 2.2 – 2.9% per million year rate of evolution (Torrioni et al 1994c).

Haplotype diversity for world populations was estimated using Equation 8.5 of Nei (1987). This equation is: $h = n (1 - \sum x_i^2) / (n - 1)$. The variance of h was calculated using a modified version of the formula given in Nei (1978) due to the haploid nature of the mitochondrial system. The modified formula is:

$$V_s(h) = 1 / (n (n-1)) * (\sum x_i^2 - (\sum x_i^2)^2 + 2(n - 1) * (\sum x_i^3 - (\sum x_i^2)^2))$$

(M. F. Hammer, personal communication)

Results

Similarities between Indian and World Populations

One hundred and eighty-seven Indian DNA samples were amplified and scored for the chosen RFLP sites (see Table 6). All RFLP sites were polymorphic in the Indian dataset. RFLP results from this sample appear in Appendix B, while the complete sorted dataset including samples from the literature appears in Appendix C. The 41 haplotypes present in the total dataset are shown in Table 8, while the dataset itself is presented in graphical form in Figure 6. All data composing Figure 6 is from published results (see Table 7) except for the Indian sample which is from the present study.

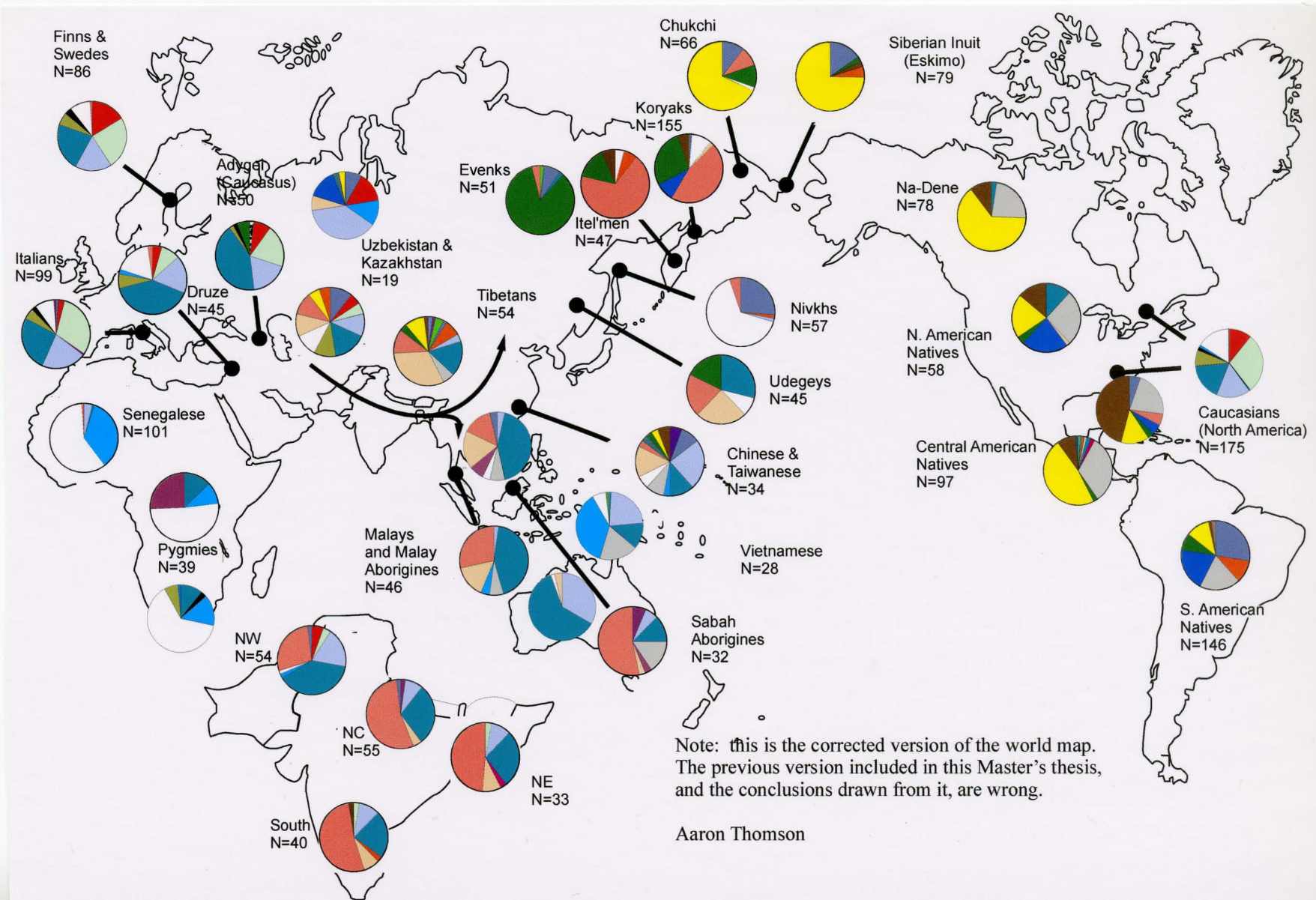
Fitch-Margoliash distance trees for the main world populations and India are shown in Figure 7. Different topologies created by including Siberia as one population or several (the Evenks, Nivkhs and Udegeys) are shown, as well as those created by different distance measurements (the Cavalli-Sforza and Edwards (1967) chord distance and Reynolds, Weir and Cockerham (1983) distance). According to Phylip 3.57c (Felsenstein, 1995) (Genedist) documentation, these are alternative estimates of the same quantity in identical models. These trees are consistently generated by the Fitch-Margoliash algorithm (Fitch and Margoliash, 1967) under conditions of randomized population entry order and global branch rearrangement (removal and re-addition of branches after the initial tree generation as a method to find any possible better trees). With the trees generated from the Cavalli-Sforza chord distance matrices (Figures 7A and 7B), it can be seen that India is more similar to European populations than Asian

Table 8: World Haplotypes. 1 indicates a restriction site or deletion presence, 0 indicates a restriction site or deletion absence.

Haplotype	663 HaeIII	5176 AluI	7025 AluI	10394 DdeI	10397 AluI	13262 AluI	16065 HinfI	16517 HaeIII	9-bp del
1	0	0	0	1	1	1	1	0	0
2	0	0	1	0	0	0	1	1	0
3	0	0	1	0	0	1	1	0	0
4	0	0	1	0	0	1	1	1	0
5	0	0	1	0	0	1	1	1	1
6	0	0	1	1	1	0	1	0	0
7	0	0	1	1	1	1	1	0	0
8	0	0	1	1	1	1	1	1	0
9	0	1	0	0	0	0	1	0	0
10	0	1	0	0	0	0	1	1	0
11	0	1	0	0	0	1	1	1	0
12	0	1	0	1	0	0	1	0	0
13	0	1	1	0	0	0	0	1	0
14	0	1	1	0	0	0	1	0	0
15	0	1	1	0	0	0	1	1	0
16	0	1	1	0	0	1	1	0	0
17	0	1	1	0	0	1	1	0	1
18	0	1	1	0	0	1	1	1	0
19	0	1	1	0	0	1	1	1	1
20	0	1	1	1	0	0	0	0	0
21	0	1	1	1	0	0	0	1	0
22	0	1	1	1	0	0	1	0	0
23	0	1	1	1	0	0	1	1	0
24	0	1	1	1	0	0	1	1	1
25	0	1	1	1	0	1	1	0	0
26	0	1	1	1	0	1	1	1	0
27	0	1	1	1	0	1	1	1	1
28	0	1	1	1	1	0	0	1	0
29	0	1	1	1	1	0	1	0	0
30	0	1	1	1	1	0	1	1	0
31	0	1	1	1	1	0	1	1	1
32	0	1	1	1	1	1	0	0	0
33	0	1	1	1	1	1	1	0	0
34	0	1	1	1	1	1	1	0	1
35	0	1	1	1	1	1	1	1	0
36	0	1	1	1	1	1	1	1	1
37	1	1	1	0	0	0	1	1	0
38	1	1	1	0	0	1	1	0	0
39	1	1	1	0	0	1	1	0	1
40	1	1	1	0	0	1	1	1	0
41	1	1	1	1	0	1	1	1	0

Figure 6: Haplotype frequencies for global populations examined.

Numbers are haplotype labels as shown in Table 8.



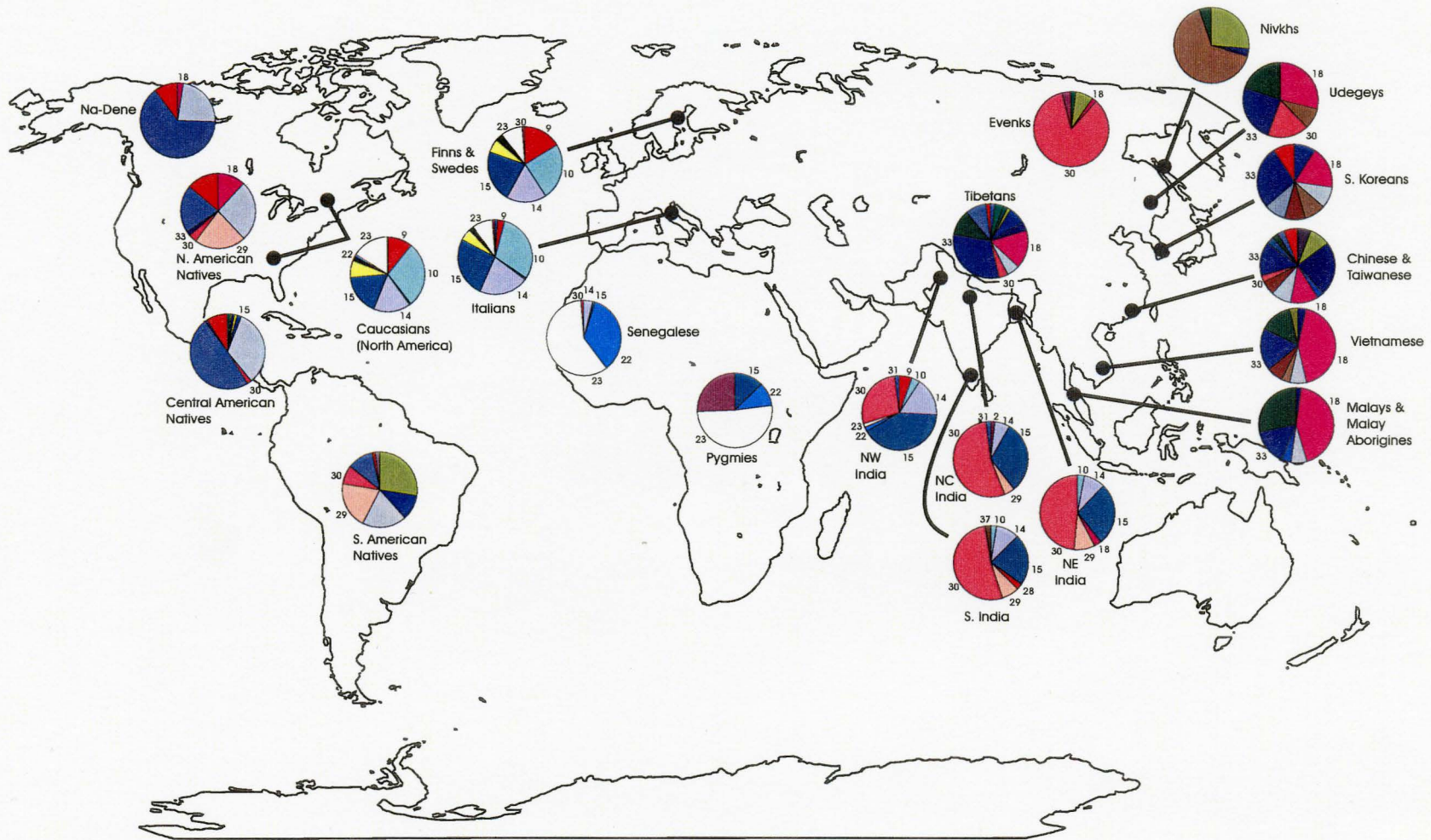
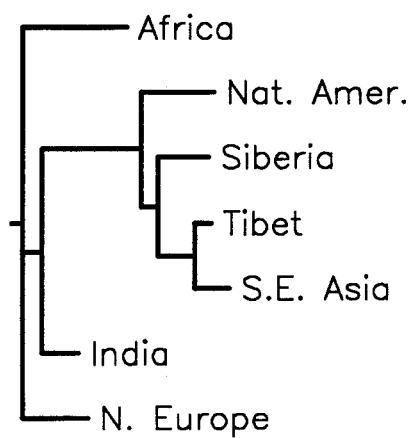


Figure 7: Fitch-Margoliash distance trees of India and world populations.

- 7A: Cavalli-Sforza chord distance matrix,
Siberia included as one population.
- 7B: Cavalli-Sforza chord distance matrix,
Siberia included as three subpopulations.

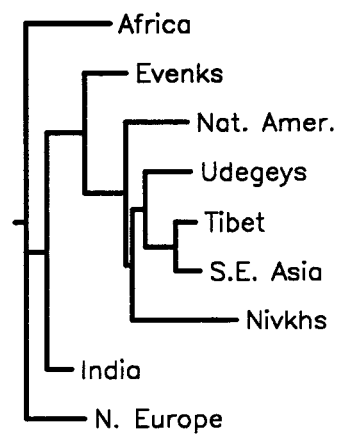
- 7C: Reynolds distance matrix,
Siberia included as one population.
- 7D: Reynolds distance matrix,
Siberia included as three subpopulations.

A



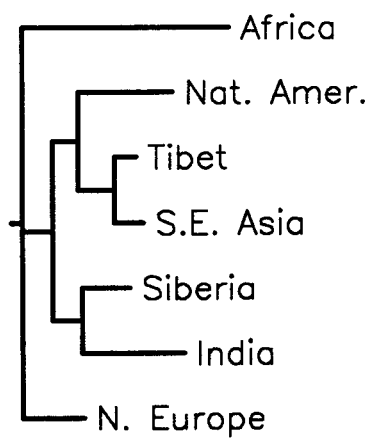
0.01

B



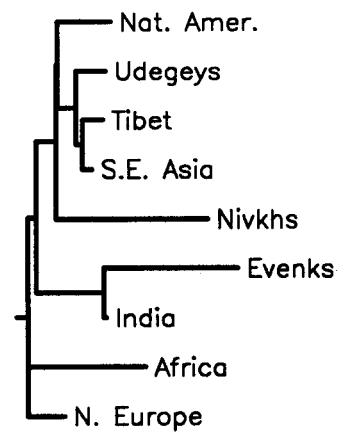
0.01

C



0.05

D



0.05

populations. When the Reynolds distance matrix is used, Siberia (Figure 7C) or the Siberian subgroup the Evenks (Figure 7D) group with India, rather than more closely to East Asian and Native American populations as in Figures 7A and 7B. It is not clear why the Reynolds distance measure gives slightly different answers than the Cavalli-Sforza chord measure with respect to Siberian populations, but any difference in handling extremely divergent allele frequencies (such as the high frequency of haplotype 30 in the Evenk population) could be responsible for the discrepancy.

Results from the Fitch-Margoliash tree generation algorithm for Figure 7 were checked with the Phylip neighbour-joining program Neighbor. Comparisons between the results of these algorithms were thought to be important due to the different strategies utilized. The Fitch-Margoliash algorithm joins pairs of operational taxonomic units separated by the smallest mutation distance until the entire group is within one set. The tree is then tested for the “standard deviation” (Fitch and Margoliash, 1967) of the mutation distances reconstructed from the tree as compared to those used as input. The tree with the smallest “standard deviation” is chosen as the “correct” tree. By comparison, the principle of the neighbour-joining algorithm “is to find pairs of operational taxonomic units (OTUs [=neighbors]) that minimize the total branch length at each stage of clustering of OTUs starting with a starlike tree” (Saitou and Nei, 1987). Conclusions with respect to the relation of India to world populations did not change as a result of the comparison, but some divergent populations (e.g. Africa and Native Americans) grouped differently for some distance matrices than with the Fitch-Margoliash algorithm. Due to the small differences between the Fitch-Margoliash and

neighbour-joining results, it was not thought necessary to include the neighbour-joining trees.

Figure 8 shows a minimum length majority-rules maximum parsimony consensus tree of world haplotypes with percentage resampling values. This tree was generated from 602 random order re-entries of the haplotypes. Due to their high resampling values, the two main haplotype clusters (designated A and B) were used as categories in all statistical tests in which haplotypes of both groups were present. In cases where haplotypes of both groups were not present in the populations being tested, 2-3 haplotype categories (dependent on sample size) were created based on haplotype order in Group A of Figure 8.

Group A and Group B of Figure 8 are separated due to the fact that all haplotypes of Group B are 663 *Hae*III (-), 10394 *Dde*I (-), 10397 *Alu*I (-), 13262 *Alu*I (-). In Group A, those haplotypes which are closest to the A-B split (3-5, 11 and 16-19) are usually separated from Group B only by the presence of the 13262 *Alu*I site. Haplotype 37 is excluded from Group B by the presence of the 663 *Hae*III (+) site. All other haplotypes in Group A are excluded from Group B by at least two restriction sites. Unfortunately, resampling values within Group A were too low to clearly differentiate sub-groups.

Figure 9 shows the consensus tree with dots indicating the presence of each haplotype with respect to the main world population locations and India. It should be noted that while haplotype 30 is present in all main world populations, it is present in low frequencies in African, European and East Asian populations (with only 1-2 samples of haplotype 30 from each of these populations). This seems to imply that haplotype 30

Figure 8: Minimum length majority-rules maximum parsimony
consensus tree of world haplotypes
with percentage resampling values.

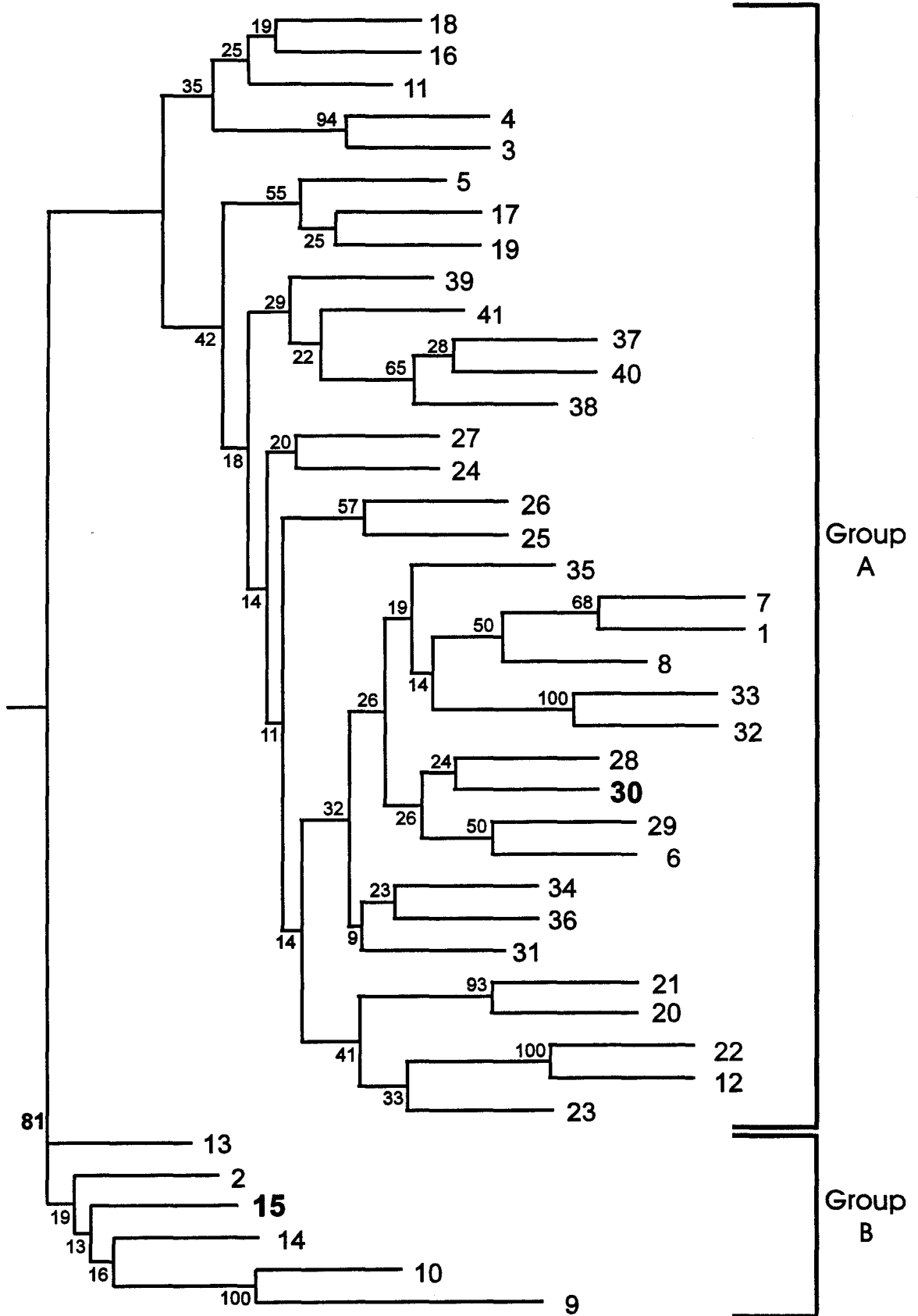
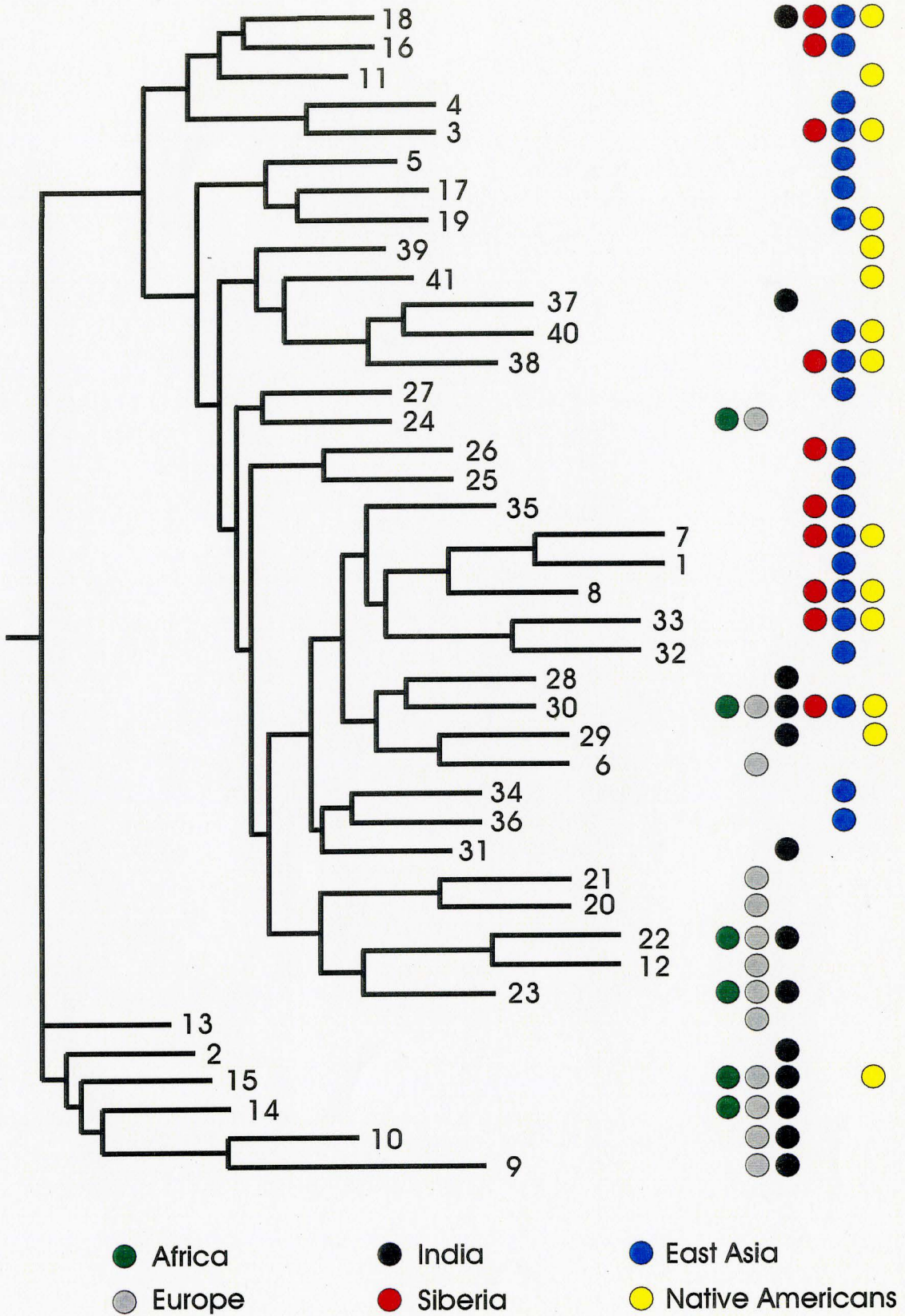


Figure 9: Maximum parsimony consensus tree
with world haplotype distributions.



likely originated in either the Siberian or Indian population. The implications of this figure will be considered in the discussion.

Statistical Tests

Various statistical tests were conducted to determine the validity of observed patterns between populations. Due to the high resampling value between Group A and B (see Figure 8), this was used as the basis for Group A vs. Group B haplotype comparisons in all populations in which haplotypes of both groups were present. When comparing populations in which Group B haplotypes were not present, Group A was broken into sub-groups based on Figure 8. Table 9 shows statistical comparisons between selected world populations and those parts of India which are closest (geographically and genetically). These comparisons show statistically significant heterogeneity; therefore all other world populations must also be significantly different from Indian populations.

Table 10 shows the calculated mitochondrial haplotype diversities for the populations in this study. It is clear that diversity calculations using RFLPs randomly distributed over the entire mitochondrial genome gives a substantially different answer (usually showing higher diversity) with respect to which populations are most diverse. The African and Evenk populations particularly show this effect. This matter will be examined in the discussion.

Table 11 contains the results of Group A vs. Groups B geographic homogeneity testing for European and African populations. These results demonstrate the validity of including only one European (caucasian) sample in the distance analysis, as the European

Table 9: Contingency Table Analysis of Haplotype Frequency Distributions between India and Nearest External Populations

Populations	Statistic				
	G-Test	X ² -Test	Monte Carlo Test	Monte Carlo 95% C.I.	Fisher's Exact Test
NW India vs. Italy ^a	G = 4.329 df = 1 *P < 0.05	X ² = 4.458 df = 1 *P < 0.05	----	----	P = 0.046
NE India vs. Thailand ^a	G = 42.35 df = 2 **P < 0.005	X ² = 33.14 df = 2 **P < 0.005	P = 0.000	0.000-9.6*10 ⁻⁵	----
NE India vs. Vietnam ^a	G = 33.44 df = 2 **P < 0.005	X ² = 26.14 df = 2 **P < 0.005	P = 0.000	0.000-9.6*10 ⁻⁵	----

* Significant P < 0.05 ** Significant P < 0.01

^a Significant when Sequential Bonferroni Correction was used on the P value from Fisher's Exact Test

Table 10: Mitochondrial haplotype diversities of world populations.

Population	n	This Study's Markers		Complete Mitochondrial Genome (RFLPs)	
		<i>h</i>	95% Confidence Interval	<i>h</i>	95% C. I.
N India	142	0.685	±0.002		
NW India	54	0.737	±0.004		
NC India	48	0.615	±0.005		
NE India	33	0.693	±0.008		
S India	40	0.673	±0.007		
NC, NE & S India	121	0.648	±0.002		
Northern Europe	86	0.831	±0.001	0.966	±0.001
Caucasians	175	0.823	±0.001	0.9787	±0.0003
Italians	99	0.790	±0.001	0.955	±0.001
Senegalese	101	0.530	±0.002	0.974	±0.006
Pygmies	39	0.661	±0.006	0.930	±0.014
Africa	140	0.592	±0.002	0.982	±0.003
Tibetans	54	0.854	±0.003	0.987	±0.001
Nivkhs	57	0.515	±0.006	0.756	±0.004
Evenks	51	0.286	±0.008	0.889	±0.002
Udegeys	45	0.795	±0.002	0.843	±0.003
Siberia	153	0.787	±0.001	0.930	±0.005
Koreans	13	0.927	±0.015	0.987	±0.007
Taiwanese Han	20	0.932	±0.005	0.990	±0.003
Malaysian Chinese	14	0.813	±0.018	0.989	±0.006
China and Taiwan	34	0.906	±0.003	0.989	±0.001
Vietnamese	28	0.778	±0.009	0.992	±0.002
Malay Aborigines	32	0.683	±0.008	0.925	±0.004
Malays	14	0.758	±0.016	1.000	- 0.005
Thailand Total	46	0.717	±0.005	0.964	±0.002
Na-Dene	78	0.531	±0.004	0.766	±0.003
North American Natives	58	0.819	±0.002	0.946	±0.001
Central American Natives	97	0.653	±0.002	0.879	±0.001
South American Natives	146	0.823	±0.001	0.946	±0.001

Table 11: Pairwise Contingency Table Analysis of Haplotype Frequency Distributions between European and African Populations

Populations	Statistic		
	G-Test	X ² -Test	Fisher's Exact Test
N. American vs. N. Europe Caucasian	G = 1.607 df = 1 P > 0.05	X ² = 1.566 df = 1 P > 0.05	P = 0.226
N. American vs. Italy Caucasian	G = 2.707 df = 1 P > 0.05	X ² = 2.630 df = 1 P > 0.05	P = 0.139
N. Europe vs. Italy	G = 0.075 df = 1 P > 0.05	X ² = 0.075 df = 1 P > 0.05	P = 0.855
Italy vs. Pygmies	G = 59.09 df = 1 **P < 0.005	X ² = 56.72 df = 1 **P < 0.005	P = 0.000
Pygmies vs. Senegalese	G = 2.373 df = 1 P > 0.05	X ² = 2.627 df = 1 P > 0.05	P = 0.141

* Significant P < 0.05 * Significant P < 0.005

populations are not significantly different. In spite of the two African populations sampled not appearing to be significantly heterogeneous, the entire African dataset was included in the distance analysis due to the highly varying frequencies of haplotypes 14, 15 and 24 within African subpopulations. All other populations and individuals found in the literature were included in distance analyses. Table 12 shows the results of statistical testing within Siberian populations, and clarifies why the Siberian sub-populations were entered into the distance analysis separately.

Table 13 shows the ages of several haplotypes and haplotype pairs as estimated from the coalescence time of 10 samples of whole-mitochondrial genome restriction data obtained from the literature. In cases where one haplotype appeared to be responsible for a large proportion of the weighted average, the analysis was checked by removing that (most divergent) sample. This indicated a case of possible parallel mutation in the haplotype-defining sites in the case of haplotype 33.

Populations Within India

Haplotypes present in the Indian sample are shown in Table 14, while haplotype frequencies within India are shown in Table 15. Cases where possible source populations of migration into India can be identified are highlighted with colour, showing (as in Figure 6) that north-western India seems to have received more migrants from Europe than has the rest of India.

Fitch-Margoliash distance trees for the main world populations and significantly different Indian populations (NW India vs. NC, NE and South India) are shown in Figure

Table 12: Pairwise Contingency Table Analysis of Haplotype Frequency Distributions between Siberian and Korean Populations, and Within Siberia

Populations	Statistic				
	G-Test	X ² -Test	Monte Carlo Test	Monte Carlo 95% C.I.	Fisher's Exact Test
Udegeys vs. Koreans	G = 1.067 df = 1 P > 0.05	X ² = 1.112 df = 1 P > 0.05	----	----	P = 0.305
Evenks vs. Koreans ^a	G = 11.58 df = 1 **P < 0.005	X ² = 14.59 df = 1 **P < 0.005	----	----	P = 0.001
Nivkhs vs. Koreans	G = 0.577 df = 1 P > 0.05	X ² = 0.593 df = 1 P > 0.05	----	----	P = 0.500
Nivkhs vs. Udegeys ^a	G = 44.81 df = 3 **P < 0.005	X ² = 39.24 df = 3 **P < 0.05	P = 0.000	0.000-9.6*10 ⁻⁵	----
Evenks vs. Udegeys ^a	G = 8.654 df = 1 **P < 0.005	X ² = 14.39 df = 1 **P < 0.005	----	----	P = 0.001
Nivkhs vs. Evenks ^a	G = 42.40 df = 1 **P < 0.005	X ² = 40.67 df = 1 **P < 0.005	----	----	P = 0.000

* Significant P < 0.05

** Significant P < 0.005

^a Significant when Sequential Bonferroni Correction was used on the P value from Fisher's Exact Test

Table 13: Haplotype coalescence time as estimated from world populations present in literature data.

Haplotype(s)	most common in	mean divergence (π)	time estimate (2.2-2.9 % per Myr rate of mtDNA evol.)	average time estimate (years)
22/23 ¹	Africa/Europe	0.0028	130,000-98,000	110,000
14/15 ¹	Europe/India	0.0025	110,000-85,000	97,000
14/15 ¹ (Excluding Chen62) ²	Europe/India	0.0020	90,000-69,000	78,000
29/30 ¹	India	0.0014	61,000-46,000	53,000
18	East Asia	0.0011	50,000-38,000	43,000
33	East Asia	0.0030	140,000-100,000	120,000
33 (excluding TorT149) ²	East Asia	0.0015	66,000-51,000	57,000

¹As haplotype pairs are separated only by the hypervariable site 16,517 *Hae* III, samples from these pairs were treated as being from the same haplotype for the purposes of this analysis.

²Samples which appeared to be responsible for a large part of the divergence observed for a haplotype were excluded from a subsequent analysis to determine their contribution to the estimated age of the coalescent.

Table 14: Indian Haplotypes

Haplotype	663 HaeIII	5176 AluI	7025 AluI	10394 DdeI	10397 AluI	13262 AluI	16065 HinfI	16517 HaeIII	9-bp del	Shared With?
2	0	0	1	0	0	0	1	1	0	unique to India
9	0	1	0	0	0	0	1	0	0	Europe
10	0	1	0	0	0	0	1	1	0	Europe
14	0	1	1	0	0	0	1	0	0	Africa, Europe
15	0	1	1	0	0	0	1	1	0	Africa, Europe
18	0	1	1	0	0	1	1	1	0	East Asia, Native Americans
22	0	1	1	1	0	0	0	1	0	Africa, Caucasians
23	0	1	1	1	0	0	1	1	0	Africa, Europe
28	0	1	1	1	1	0	0	1	0	unique to India
29	0	1	1	1	1	0	1	0	0	Native Americans
30	0	1	1	1	1	0	1	1	0	Africa, Europe, East Asia, Native Americans
31	0	1	1	1	1	0	1	1	1	unique to India
37	1	1	1	0	0	0	1	1	0	unique to India

Table 15: Indian Haplotype Frequencies

Colours are for clarification only and have no connection to Figures 6 or 11.

	Shared with other populations								Unique to India				
Haplotypes	9	10	14	15	18	22	23	29	30	2	28	31	37
NW India	5.6	3.7	16.7	40.7	0.0	1.9	1.9	0.0	27.8	0.0	0.0	1.9	0.0
NC India	0.0	0.0	6.1	32.7	0.0	0.0	0.0	4.1	53.1	2.0	0.0	2.0	0.0
NE India	0.0	3.0	9.1	27.3	3.0	0.0	0.0	9.1	48.5	0.0	0.0	0.0	0.0
S India	0.0	2.5	10.0	22.5	0.0	0.0	0.0	7.5	52.5	0.0	2.5	0.0	2.5

Possible evidence of gene flow into India from:

- Europe
- East Asia
- Africa
- Siberia

10. All individuals from areas (not including Maharashtra - see Figure 5) were included in these analyses, regardless of religious or caste affiliation. It can be seen that Indian populations are most similar to other Indian populations. However, when compared individually to world populations, north-western India is more similar to European populations, while north-central, north-east and southern India share a strong similarity to Siberian populations. These differences are due mainly to haplotype frequency differences, with the differential presence or absence of haplotypes from geographic areas making a smaller contribution. As before, comparison of the Fitch-Margoliash results with the neighbour-joining results did not alter the conclusions.

A comparison of haplotypes between religions and caste groups in the Indian sample is shown in Figure 11. It can be seen that in general, the different religions of India appear to be quite different genetically (in fact, the Muslim and Christian samples are significantly different), while the Hindu castes appear to be quite similar. However, these initial impressions may not be completely accurate, as will be discussed later.

Figure 12 shows the distribution of haplotypes for various religions in the Indian subcontinent, while Figure 13 shows the distribution of haplotypes among Hindu castes, Hindus of unknown caste and Hindu-derived religions with respect to our sample.

Statistical Tests

Table 16 shows the results of all Group A vs. Group B geographical statistical comparisons within India which were significant, while Table 17 shows the results of all similar statistical comparisons within India which were not significant. Northwestern

Figure 10: Fitch-Margoliash distance trees of Indian subsections and world populations.

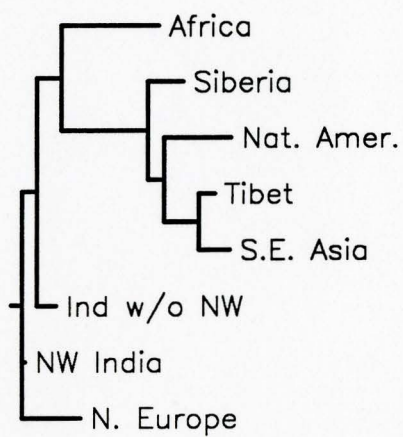
10A: Cavalli-Sforza chord distance matrix,
Siberia included as one population.

10B: Cavalli-Sforza chord distance matrix,
Siberia included as three subpopulations.

10C: Reynolds distance matrix,
Siberia included as one population.

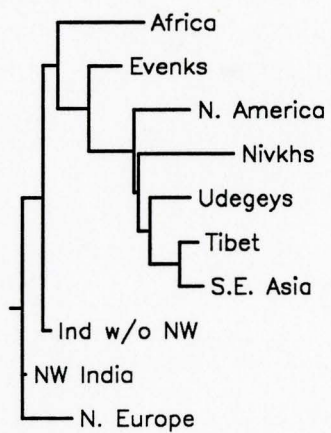
10D: Reynolds distance matrix,
Siberia included as three subpopulations.

A



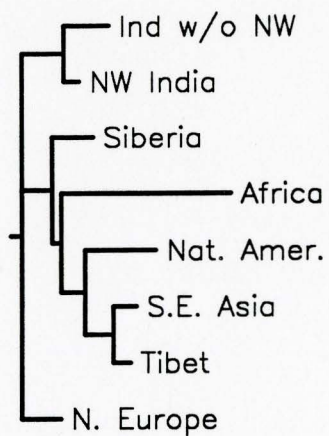
0.01

B



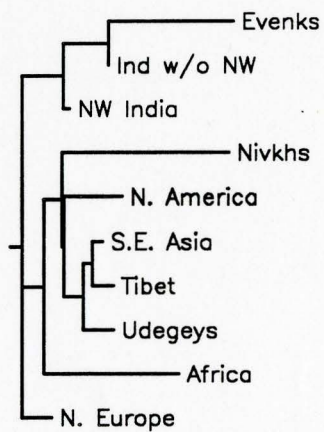
0.01

C





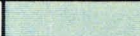
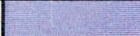



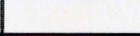





0.05

D

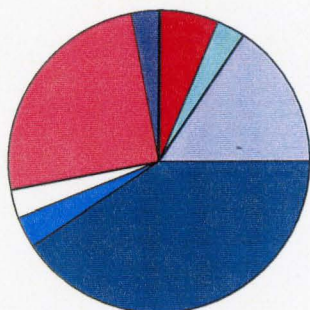


0.05

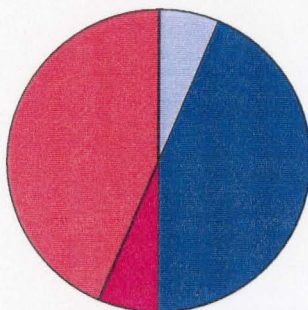
Figure 11: Comparison of haplotype frequencies within religious groups and caste groups.
Haplotype colours are identical to those in Figure 6.

Haplotype Legend	
2	
9	
10	
14	
15	
18	
22	
23	
28	
29	
30	
31	
37	

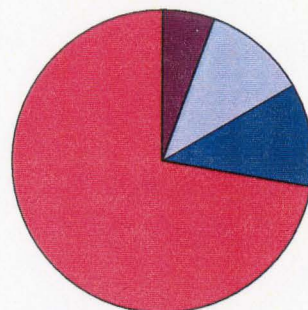
Haplotype Distribution Within Religious Groups



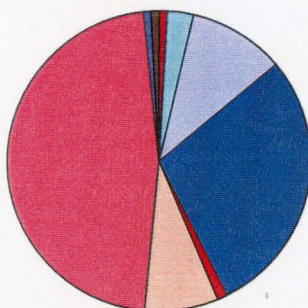
Sikh
n=30



Muslim
n=16

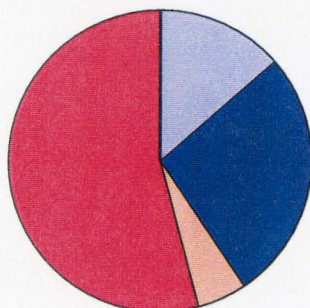


Christian
n=18

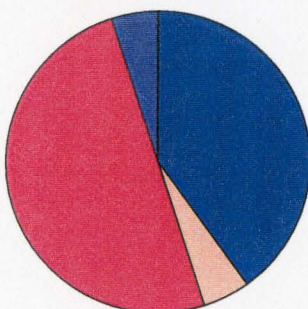


Hindu
n=115

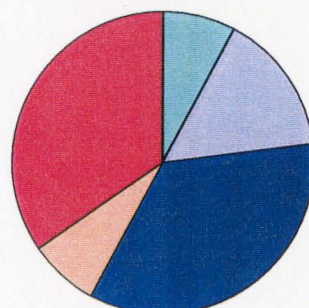
Haplotype Distribution Within Castes



Brahmin
n=37

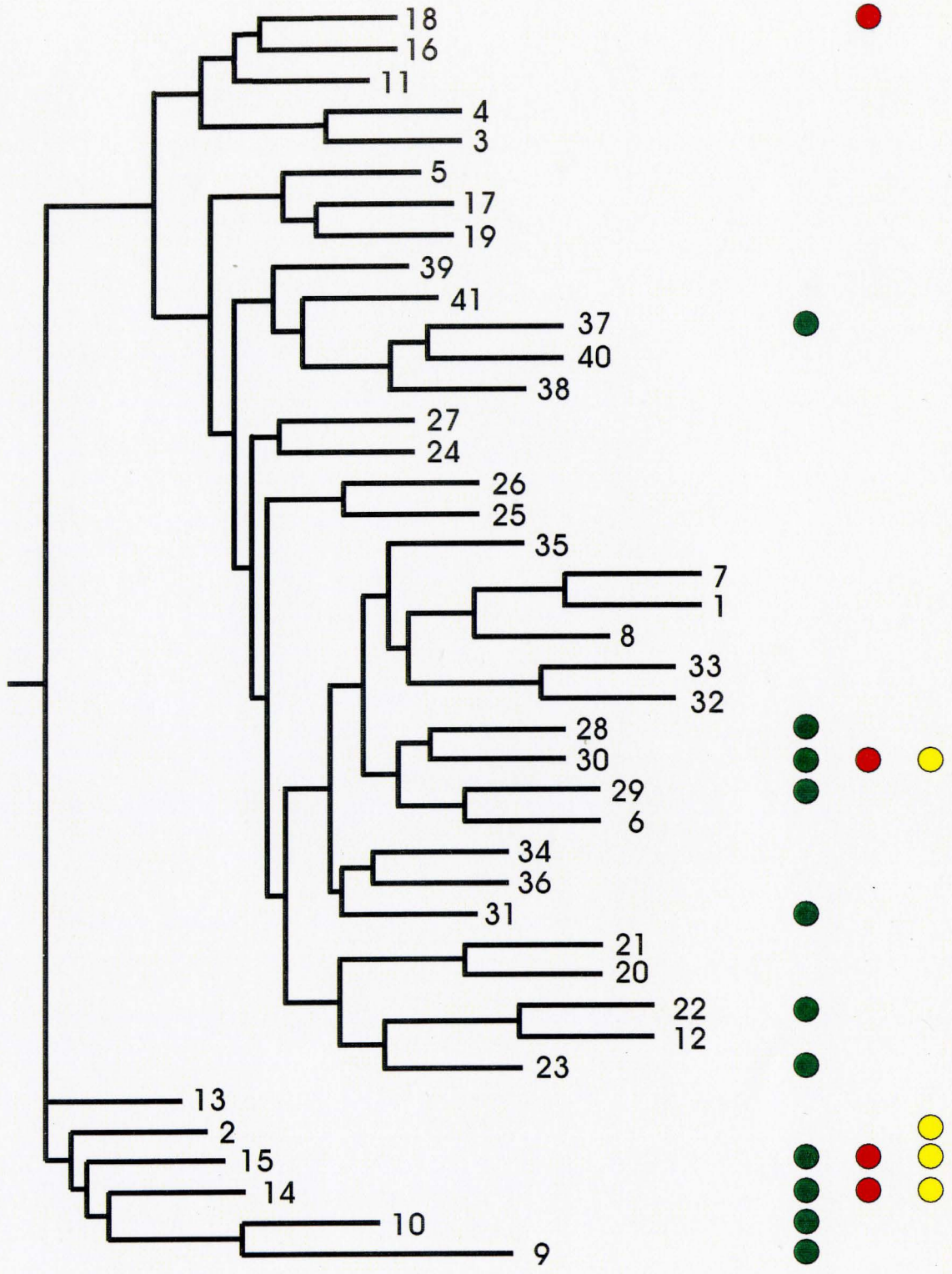


Ksatriya
n=21



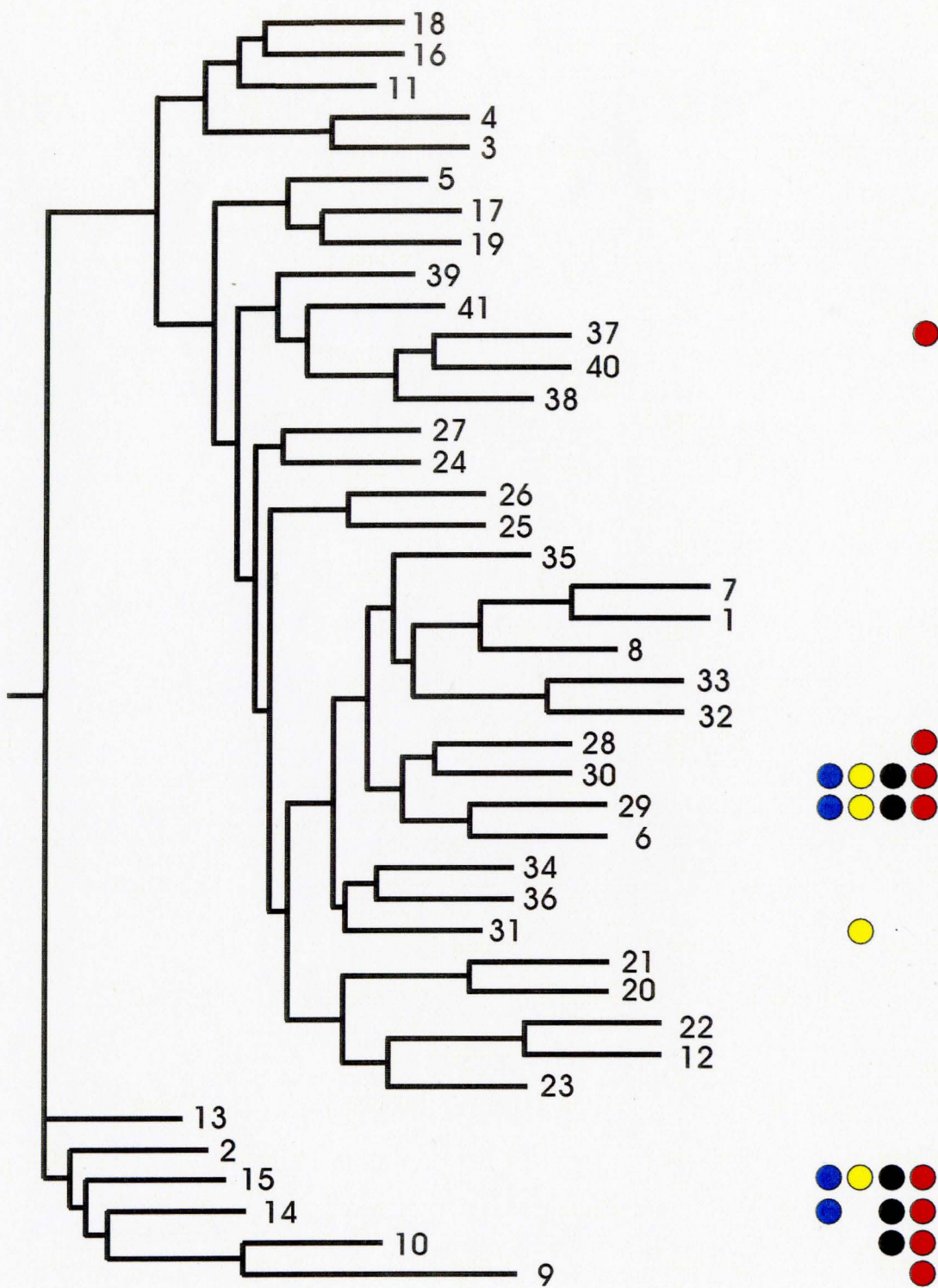
Vaisya
n=29

Figure 12: Maximum parsimony consensus tree
with Indian religious distributions.



● Hindu, Sikh and Jain ● Muslim ● Christian

Figure 13: Maximum parsimony consensus tree
with Indian caste distributions.



● Brahmin ● Kshatriya ● Vaisya ● Unknown

Table 16: Contingency Table Analysis of Haplotype Frequency Distributions Between areas Within India – **Significant Tests**

Populations	Statistic		
	G-Test	X ² -Test	Fisher's Exact Test
NW India vs. S India ^a	G = 9.389 df = 1 **P < 0.005	X ² = 9.255 df = 1 **P < 0.005	P = 0.003
NW India vs. NC India	G = 7.587 df = 1 **P < 0.01	X ² = 7.502 df = 1 **P < 0.01	P = 0.0095
NW India vs. NE India	G = 6.218 df = 1 *P < 0.025	X ² = 6.193 df = 1 *P < 0.025	P = 0.016
NW India vs. S India (non-Sikh) (non-Christian)	G = 5.521 df = 1 *P < 0.025	X ² = 5.408 df = 1 *P < 0.025	P = 0.025

* Significant P < 0.05 ** Significant P < 0.01

^a Significant when Sequential Bonferroni Correction was used on the P value from Fisher's Exact Test

Table 17: Contingency Table Analysis of Haplotype Frequency Distributions Between areas Within India – Non-Significant Tests

Populations	Statistic		
	G-Test	X ² -Test	Fisher's Exact Test
NC India vs. NE India	G = 0.0003 df = 1 P > 0.05	X ² = 0.0003 df = 1 P > 0.05	P = 1.000
NC India vs. S India	G = 0.196 df = 1 P > 0.05	X ² = 0.196 df = 1 P > 0.05	P = 0.828
NE India vs. S India	G = 0.150 df = 1 P > 0.05	X ² = 0.150 df = 1 P > 0.05	P = 0.809
N India vs. S India	G = 1.887 df = 1 P > 0.05	X ² = 1.858 df = 1 P > 0.05	P = 0.209
SW India vs. SE India (Kerala & Karnataka) (Tamil Nadu & Andhra Pradesh)	G = 1.283 df = 1 P > 0.05	X ² = 1.283 df = 1 P > 0.05	P = 0.327
NW Sikh vs NW non-Sikh	G = 0.152 df = 1 P > 0.05	X ² = 0.151 df = 1 P > 0.05	P = 0.776
S. Christian vs S. Non-Christian	G = 0.029 df = 1 P > 0.05	X ² = 0.029 df = 1 P > 0.05	P = 1.000

* Significant P < 0.05 ** Significant P < 0.01

^a Significant when Sequential Bonferroni Correction was used on the P value from Fisher's Exact Test

India is significantly different from north-central, north-east and southern India, but north-central, north-east and southern India are not significantly different when compared with each other. However, when the Bonerroni correction was done on this data, only the Northwestern India vs. South India test was still significant. This appears to be due to the low frequency of haplotypes 10, 14 and 15 in South India, along with the higher frequency of haplotypes 29 and 30. Due to the large numbers of Sikhs in the northwestern sample and Christians in the southern sample, tests for heterogeneity between religious groups within these geographical area were necessary. These tests did not indicate a significant difference between the minority religious groups and the majority of the population, as expected by the history of these groups (conversion from the majority Hindu population). Unfortunately, sample size was too small to expect to be able to discern any small differences between the Hindu majority and minority populations (excluding differences due to geographic sampling).

Table 10 shows the calculated mitochondrial haplotype diversities for the populations in this study. It is obvious that north-western India is significantly more diverse than the rest of the subcontinent. Unfortunately, as whole-genome restriction data was not available for this population, it was not possible to determine the unbiased haplotype diversity.

Table 18 shows the results of Group A vs. Group B comparisons between castes and various religions present within India. Most comparisons are non-significant, demonstrating homogeneity between groups on the subcontinent geographical scale. Those comparisons which show heterogeneity between populations before the Bonferroni

Table 18: Contingency Table Analysis of Haplotype Frequency Distributions between Religions Present Within India

Populations	Statistic		
	G-Test	X ² -Test	Fisher's Exact Test
Brahmin vs. Kshatriya	G = 0.002 df = 1 P > 0.05	X ² = 0.002 df = 1 P > 0.05	P = 1.000
Brahmin vs. Vaisya	G = 1.807 df = 1 P > 0.05	X ² = 1.801 df = 1 P > 0.05	P = 0.208
Kshatriya vs. Vaisya	G = 1.423 df = 1 P > 0.05	X ² = 1.415 df = 1 P > 0.05	P = 0.373
Brahmin vs. Sikh	G = 4.379 df = 1 *P < 0.05	X ² = 4.327 df = 1 *P < 0.05	P = 0.054
Kshatriya vs. Sikh	G = 3.290 df = 1 P > 0.05	X ² = 3.276 df = 1 P > 0.05	P = 0.090
Vaisya vs. Sikh	G = 0.383 df = 1 P > 0.05	X ² = 0.383 df = 1 P > 0.05	P = 0.594
Brahmin vs. Christian	G = 0.872 df = 1 P > 0.05	X ² = 0.852 df = 1 P > 0.05	P = 0.391
Kshatriya vs. Christian	G = 0.633 df = 1 P > 0.05	X ² = 0.629 df = 1 P > 0.05	P = 0.506

Vaisya vs. Christian	G = 3.937 df = 1 *P < 0.05	X ² = 3.839 df = 1 P > 0.05	P = 0.069
Sikh vs. Christian	G = 6.781 df = 1 *P < 0.05	X ² = 6.611 df = 1 *P < 0.05	P = 0.018
Muslim vs. Christian	G = 1.783 df = 1 P > 0.05	X ² = 1.771 df = 1 P > 0.05	P = 0.291

* Significant P < 0.05

Correction – Brahmin vs. Sikh, Vaisya vs. Christian and Sikh vs. Christian – will be examined in detail in the discussion.

Discussion

India and other World Populations

It is clear from Figures 6, 7 and 10 that the populations with the most similarity to those in India are not the Tibetans, Vietnamese or Malays. Rather, the European and Siberian (Evenk) populations in the dataset appear to be related most closely to the groups which have made the largest contributions to the mitochondrial gene pool of India. This is consistent with previous linguistic, classical marker and mitochondrial studies (Cavalli-Sforza *et al*, 1994; Roychoudhury and Ney, 1985; Barnabas *et al*, 1996) which have shown similarities between European and Indian populations, and also the main Y-chromosomal study (Hammer *et al*, 1998) which showed haplotype similarities with central (and northern) Asian populations. This is also consistent with the standard historical model for the origin of contemporary Indian populations – that at least two separate migrations appear to have contributed to the present-day population of the subcontinent. If the Indo-Aryans had originated as a minor ethnic group within the subcontinent, as some authors have recently promulgated (see Agrawal, 1996 and Wallia, 1996 for reviews of some recently published books), then this pattern would likely not be present due to similarities between the Indo-Aryans and their neighbors. The distribution of haplotypes 29 and 30 is definitely not consistent with any proposed migration from India to Europe, which is an obvious corollary to the theory of Indo-Aryan origin within India (due to the undisputed similarities between Sanskrit and European languages).

The Hammer (1998) paper showed that specifically, the Mongolian population (followed by the West Siberians) displays the most similarity to South Indians with respect to the Y-chromosomal gene pool. There is no whole-mitochondria restriction data available for Mongolia, so it was not possible to directly determine whether Mongolian populations are more similar to Indians than the Evenks for this marker. However, some indirect mitochondrial evidence points to a possible relationship between Indians and Mongolians. Mitochondrial DNA RFLP and control region analysis has shown that the Mongolian population has the most similarity to Native Americans of all Asian populations (Kolman *et al.*, 1996; Merriwether *et al.*, 1996), while haplotype 29 in the dataset exists only in Indians and Native Americans. Other haplotypes are shared between Indians and Native Americans, but these also exist in north-east Asian populations (i.e. the Evenks, Udegeys and S. Koreans) and could have been contributed to the Native American genepool by populations in north-east Asia. Given the Mongolian-Native American similarity, it seems possible that haplotype 29 is present in this population, notwithstanding its absence elsewhere in East Asia.

Unfortunately, our Indian mitochondrial dataset does not include tribal samples, so that the origin of these possibly most ancient inhabitants of the subcontinent remains uncertain. However, as tribal populations compose only 7% of the population of India (Cavalli-Sforza *et al.*, 1994), this should not alter the conclusions about the origin of the majority of Indians.

When one examines the haplotypes that are distributed throughout India and compares them with world populations, it is clear that haplotypes 29 and 30 are found

mainly in East Asia and Native Americans but haplotypes 10, 14 and 15 are associated with Europeans (Figure 6 and Table 15). This led to the intuitive conjecture that the Asian-associated haplotypes were brought into India by the Dravidian migration, while the European-associated haplotypes entered India with the Aryans. If this hypothesis were correct, it would appear that in north-central, north-east and southern India the Dravidians and Indo-Aryans had made approximately equal contributions to the genepool.

However, it is clear that this hypothesis is based on the unproven premise that the Aryan mitochondrial genepool was in fact significantly different from that which had existed in India previously. This premise is uncorroborated due to the ambiguous geographical origin of the Dravidians (less so the Aryans) and a lack of datapoints in the global dataset between Italy and Tibet. If the Dravidians originated from a population near (for example) the Caspian sea, they might be expected to have originally had some frequency of “European-characteristic” haplotypes due to their geographic origin near Europe. In this case, the present-day haplotype distribution observed in north-central, north-east and southern India might not be significantly different from the “original” Dravidian haplotype distribution. The lack of European-characteristic Y chromosome haplotypes and haplotype frequencies in South India (Hammer *et al*, 1998) is suggestive (but definitely not conclusive) that Indo-European genetic influence on this part of the subcontinent was minimal.

With the Dravidian inhabitants of India perhaps already having “European” haplotypes at some frequency, the influence of the Aryan migration with respect to

mitochondria might be limited to the modification of haplotype frequencies (especially in the northwest) and perhaps introducing haplotypes 9, 10, 22 and 23. In this scenario the genetic influence (as seen by the mitochondrial locus) of the Indo-Aryans would be much less than their observed cultural influence. However, very little evidence appears to exist which would allow discrimination between the effects of the Indo-Aryan takeover of the subcontinent and the earlier Dravidian migration.

On the basis of the mitochondrial haplotypes present within India, it appears that migration (from populations other than Europeans and Siberians) into India may have had a limited but observable impact in the north-west and north-east. For example, haplotypes 22 and 23 are present in the northwest, and while haplotype 23 is relatively common in European and Caucasian populations, haplotype 22 is rare (2 samples out of 360). Unfortunately the dataset does not contain any Middle Eastern populations, so it is not possible to determine the proximate source of haplotype 22 in north-western India, or whether northwest India might simply be the geographic terminus of a frequency cline for this haplotype. The ultimate origin of haplotype 22 would seem to have been in Africa, however, as haplotypes 22 and 23 make up the majority of the African population.

Haplotype 18 is present in north-eastern India, possibly as a result of migration from Southeast Asia into India. However, at this time it is not possible to distinguish between this possibility and the alternative hypothesis – that haplotype 18 in north-eastern India is a remnant of mitochondrial lineages which contributed substantially to the current populations of Southeast Asia. It has been observed that the Indian

subcontinent is situated geographically so as to have been a likely migration route between Africa and East Asia (Chu *et al* 1998). It seems possible that Indian populations before the Dravidian and Indo-Aryan migrations might have displayed much more similarity to the current South-east Asian populations, and that in north-eastern India all but a few of these these pre-existing mitochondrial lineages were replaced by subsequent migrations. Sequencing and/or whole mitochondrial genome RFLP typing in South-east Asia and North-east India will likely be necessary to determine which populations contain older lineages and which contain derived ones.

In comparison to similarities between other geographically adjacent world populations (e.g. northern Europeans and Italians, Chinese & Taiwanese and South Koreans), India shows little similarity to Tibetans (due to the Himalayas) and even less similarity to populations in Thailand and Vietnam (~1800 km away). The observed similarities with Africa and East Asia could have been caused by or antedate the Dravidian or Aryan migration (which occurred approximately 3,000 years ago). If this were so, it would imply that migration and gene flow into India in the last 3,000 years has not altered the mitochondrial gene pool on a large level.

Table 10 shows an interesting result with respect to world populations – African populations appear to have among the lowest levels of haplotype diversity of all groups sampled when examined using this study's haplotypes. This is not in accordance with previously published results (Templeton, 1997) which show Africa as having higher diversity at nuclear and mitochondrial loci. This is almost certainly due to the mitochondrial RFLPs chosen for this study – polymorphic markers were chosen in order

to allow discrimination between populations which were known or thought to have contributed to the Indian gene pool. Therefore, sites useful with respect to South-east Asian and European populations were studied preferentially. When the haplotypic diversity of the entire mitochondrial genome is examined using literature RFLP data (see Table 10), African populations can be seen to be among the most diverse on the planet, in concordance with previous results. The disparity between this study's markers and the results obtained from the restriction of the entire mitochondrial genome is particularly severe with the Evenk population. This makes it clear that haplotype 30 (which makes up the majority of the Evenk population) includes a large amount of diversity which was not detected with this study's markers. The presence of this diversity makes it clear (as in Table 13) that haplotype 30 is not new, and implies somewhat greater diversity for north-central, north-east and southern India than was seen with the markers of this study.

Some southeast Asian populations in Table 10 appear to be slightly more diverse than the African population, but this is likely at least partially due to the fact that the sample sizes for most of these populations were relatively small – of the more diverse south-east Asian groups, the China and Taiwan combined group had the largest sample size at 34. Interestingly, the Tibetan population appears to be approximately as diverse than the African population, if sample size (54) is considered. This is possibly due to historical contact with the Mongol Empire and intermittent political domination by China (Grinfeld, 1998), each of which would promote some level of migration into the area.

Figure 9 shows the geographical locations of world haplotypes as related to the maximum parsimony consensus tree. The pattern here seems to be one of geographical

assortment of haplotypes, with subsequent migration blurring the picture somewhat. For example, haplotype 30 is present in all world population groups, but is at low frequencies in African, European and most South-east Asian populations. Again, it is clear that the majority of the haplotypes present in India are those shared with European populations. All parts of India are also significantly different from neighbouring world populations (undoubtedly partly due to the mountainous terrain surrounding the northern parts of the subcontinent), and the main global population groups are significantly different from each other. Some localized population groups in the world dataset are also significantly different from each other (e.g. the Evenks Vs the Udegeys or Nivkhs). However, other populations separated by large distances (e.g. North Europeans vs. Italians) are not significantly different. Possible future investigation into how factors such as population size and migration rates affecting these groups have differed historically could yield interesting results.

The split between the two main groups A and B seen in the consensus haplotype tree (Figure 8) was used as a basis for statistical testing due to its relatively high resampling value. However, it is not clear what the historical and biological basis of this split is. Group B does not appear to have a recent origin, as the haplotype pair 14/15 has an estimated age of 78-96 thousand years. As Group B haplotypes are most common in Europe, two relatively recent events in this area which may have affected mitochondrial diversity should be kept in mind. The origin of agriculture in the Middle East almost certainly allowed one group of individuals to expand at the expense of their neighbours (Cavalli-Sforza *et al*, 1994). However, the isolation and loss of populations due to

European glaciation, with subsequent expansion of the remnant populations (Cavalli-Sforza et al, 1994; Torroni et al, 1998) could possibly also have had an effect on mitochondrial diversity in this area.

The age of haplotypes 14/15, 18, 29/30 and 33 were estimated using whole-mitochondrial genome restriction data from ten literature samples each (picked randomly in the populations in which the haplotypes were present.) This indicated that haplotype 15 is approximately 25,000 to 40,000 years older than haplotype 30, and that haplotype 18 is ~9000 years younger than haplotype 30 (Table 13). Interestingly, haplotype 33 appeared to be the oldest haplotype on the planet (117400 years) until the Tibetan sample #149 (TorT149) was excluded; this lowered the estimated haplotype age to a more moderate 57500 years. From this, it appears likely that the sample TorT149 was classified as haplotype 33 due to parallel evolution in the haplotype-defining sites. In general, the trend in haplotype ages is clear: Africa contains the oldest haplotypes (haplotype pair 22/23, at an age of approximately 110,000 years), followed by Europe and India, then East Asia (haplotypes 18 and 33, at ages of approximately 43-57 thousand years, respectively).

It is possible to view these haplotype age results as a possibly being due to of out-of-Africa migration patterns (i.e. Europe is closer to Africa, therefore Europe has possibly been inhabited longer – and has older haplotypes - than East Asia). However, it should be noted that this could partly be an artifact of the markers chosen for this study. As seen in Table 6, most of the markers used in defining haplotypes and haplotype pairs were chosen to help distinguish between Asian populations, partly due to geographical

proximity and partly due to available literature data. If more European-specific polymorphisms had been used (for example, the hypervariable *Hae* III 16,517 site was excluded), the haplotype pair 14/15 might have been split into smaller groups, each with a younger estimated age (as a smaller sample almost certainly must have a more recent coalescent than a larger sample). The fact that the majority of the African population is composed of one haplotype pair (but were shown in Table 16 to be very diverse when the entire mitochondrial genome was examined) emphasizes the fact that some bias is present due to RFLP choice.

Population Structuring Within India

It is clear from Figures 6, 7, 10 and Table 9 that the north-western subcontinent is significantly more similar genetically to European populations than is the southern area of the subcontinent. While north-central and north-east India appear to be significantly different from north-west India, P values from Fisher's Exact Test were not significantly different when the Sequential Bonferroni Correction was used. Higher sample sizes in these areas might resolve this apparent discrepancy. Significant differences between the north and south areas of the subcontinent had been noted previously with respect to the *Dde*I 10,394, *Alu*I 10,397 and *Alu*I 7,025 mitochondrial markers (Passarino *et al*, 1996), and was cited as evidence for reduced Indo-Aryan influence in the southern section of the subcontinent. However, the Passarino study had no samples from the north-central or northeast sections of the subcontinent, so results were incomplete. The Barnabas *et al* (1996) paper, which concluded that the north Indian population had a more recent

admixture of Caucasian mtDNA, also suffered from this deficiency of sampling locations. With samples from north-central and north-east areas, it is possible to see the large-scale homogeneity that exists between the north-central, north-east and southern areas of the subcontinent, and that simple “north-south” differentiation as such does not appear to exist for these mitochondrial sites. (See Table 16.) Instead, the northwest appears to be differentiated from the rest of the subcontinent. Besides lack of samples from the north-central and north-east in previous studies, this disparity in results is likely partly due to differing definitions of “north”, as the north-western area in this study included both the north (Punjab) and central (New Delhi) samples from the Passarino study (see Figure 5).

Differences in the northwest are likely mostly due to the fact that this area is on the main route of multiple invasions that have occurred into India, with some (unknown) amount of mitochondrial gene flow contributed by each (and some of the pre-existing lineages removed by each invasion). Separating out the effect of any one group of invaders (e.g. the Huns or Arabs) would likely require sequencing from Indian populations and the probable source populations. However, it is noteworthy that the longest occupation of this area by a dominant group was the Aryan period of consolidation (~2500 B.C. to ~800 B.C.) prior to the conquest of the rest of the subcontinent. It has been suggested that Indo-Aryan contributions to the mitochondrial genepool may have been “diluted” during the conquest of the rest of the subcontinent (Passarino *et al*, 1996), but it seems equally likely that some pre-Aryan populations of the northwest may have been removed while the Aryans were occupying the area.

Diversity was also measured for all sub-sections of the subcontinent. While non-random RFLP marker choices may somewhat invalidate the diversities measured for global populations, this would not seem to apply to the within India results. Differences between Indian subcontinent regions are mainly due to haplotype frequency differences, and therefore the previously stated caveats with respect to haplotype detection would seem not to apply. This noted, northwestern India appears to be significantly more diverse than the rest of India (Table 10). This is in agreement with previously reported results (Barnabas *et al*, 1996), which found North India to be more diverse than South or Central India (Maharashtra). Higher diversity within the northwest is undoubtedly due to the admixture of world lineages in this area. Unfortunately, our limited number of samples in central India preclude any comment on diversity within Maharashtra, which has been reported by Barnabas *et al* (1996) to be much lower than North or South India. Also, “north” and “south” were not defined (Barnabas *et al*, 1996), producing difficulties for attempted comparisons with historical data, census data or other studies.

Religious and Caste Differentiation Within India

Figure 11 appears to imply that religions within India are substantially different with respect to their mitochondrial gene pool. However, Figure 3 shows that there are substantial geographical factors involved with this picture – the Sikh sample is almost entirely from the northwest, while the Christian sample is mainly from the south. In fact, the Sikh sample does not differ significantly from the rest of the north-west sample (which is mainly Hindu, see Figure 3), and the Christian sample is not significantly

different from the (mainly Hindu) southern samples. This is an unsurprising result, as minority religious groups within India are mainly the result of conversion from the majority Hindu population (which appears to be implied in Figure 12 and the fact that minority groups contain subsets of haplotypes from the Hindu population). However, almost 70% of Christians in India live in the four southernmost provinces, and over 85% of Sikhs in India live in Punjab (Gupta, 1961). Therefore, Christian and Sikh mitochondrial gene pools appear to have significantly different haplotype frequencies (see Table 17), but this is a reflection of geography and religious conversion rates likely differing by area, not any intrinsic difference between the groups. Significant differences between the Sikhs and Christians were not found when the Bonferroni Correction was applied, but this is likely due to insufficient sample sizes (30 Sikhs vs 18 Christians) as the northwest previously was shown (with a larger sample size) to be significantly different from the south.

Results with respect to caste show a generally similar picture. Due to the geographically large-scale sampling process involved in this study, any possible differences between castes (as seen in Bamshad *et al*, 1998) have been overwhelmed by distance effects within each sampled caste group. It is clear that if one averages samples over the entire subcontinent, the caste groups do not appear significantly different when examined by mtDNA RFLP methods. This is consistent with the classical-marker based conclusion that “geographically clustered castes are more similar regardless of social hierarchy” (Bamshad *et al*, 1996). If large, continent-wide differences ever did exist between caste groups with respect to the mitochondrial locus, these patterns appear to

have been erased in the 100-150 generations since the caste system was imposed. The factors likely to be responsible for this would be subsequent migrations (adding variation to various caste levels), replacements (by removing castes (e.g. Ksatriya) from the population and replacements appearing later from within the local population) and gene flow between castes. Unfortunately, our sampling methods do not allow us to differentiate what happens on a local, within village (or district) level with respect to gene flow and castes. Therefore, small differences in the form of small genetic distances between local caste groups (including caste groups within the Sudra and Untouchable varnas) as reported by Bamshad *et al* (1998a) would not be visible in this study.

It has also become clear that the present caste sample is not completely independent of geographical effects. In Table 17, Brahmins appear to be significantly different from Sikhs when the data is examined using the G test and the X^2 test, but not with Fisher's exact test. However, this is due to the fact that only 6 of the 37 Brahmin samples are from the north-west (see Figure 4); therefore this is partially a geographical test. Similarly, the Vaisya sample appears to be borderline significantly different from the Christian sample (significant by the G-test, but not by the X^2 or Fisher's Exact Test.) However, the majority of Vaisya samples are from the north-west and north-east regions, which have the lowest frequencies of haplotype 30 in the subcontinent. The Christian sample (mainly from the south) has the highest haplotype 30 frequency of any Indian subgroup. Therefore, it would appear to be difficult to draw firm conclusions on relations between the major Hindu castes and minority religions from the present Indian dataset.

Conclusions

Indian populations are significantly different from all world populations analyzed, including those that are geographically closest (Tibetans and south-east Asians). The world populations that are most similar to India are the Europeans, followed by the Siberians (Evenks). Effects of probable migration into India from world populations are most visible in the northwest and north-east sections of the subcontinent, as some externally common haplotypes are present in low frequencies in these areas. These results are consistent with the standard historical model for the origins of the present culture and mitochondrial genepool of non-tribal populations of India, with at least two main in-migrations of world populations.

Due to historical in-migration, the northwest is significantly more diverse than other areas of India, and also significantly more similar to Europe and Africa than is southern India. North-central, north-east and southern areas of India are not significantly different from each other and are more similar to Evenk populations than is northwestern India, mainly due to the higher frequency of the haplotype shared between India and the Evenks. Religious groups appear to be subsets of the general population with respect to their mtDNA, and therefore are distinguished from each other only by different geographical centralization. When averaged over the entire subcontinent, caste groups were not found to be significantly different from each other.

Appendix A – Protocols, Solutions and Buffers

DNA Extraction with Phenol

Materials:

Phenol	Lysis Buffer (1X)
Chloroform iso-amyl alcohol (24:1)	Saline solution
SSTE	Autoclaved 1.5 ml tubes
NaAc (2M, pH 5.2)	EtOH (100%)
Proteinase K	TE

Whole blood preparation:

Use 1 ml whole blood in a 1.5 ml tube, and spin for @ minutes @ 8000 rpm.

Remove about 0.5 ml plasma (using the 1000 µl pipetter), leaving cellular material in the tube

Lyse red blood cells:

Add 1 ml cold *Red Cell Lysis buffer* (1X), and invert the tube several times to mix.

Let stand on ice for 15 minutes.

Spin for 2 minutes @ 8000 rpm and remove the top layer of liquid. The liquid being removed should not include cells, but will be dark red for the first few lysos. Stop where the liquid becomes cloudy (likely about 0.75 ml).

Resuspend pellet well and add 1 ml Red Cell Lysis Buffer. Spin and remove plasma.

Repeat as necessary (excluding the 15 minute wait). 3 lysos should be sufficient.

When a good white blood cell pellet is obtained (without cellular junk), wash with 300 ml saline solution (then spin @ 8000 rpm for 2 minutes & pipette off the supernatant)

Lyse white blood cells:

Add 300 μ l SSTE and 100 μ l proteinase K (20 mg/ml): total volume = 400 μ l

Resuspend pellet (liquid will be globular to gelatinous)

Incubate @ 55°C for 1 hour to overnight (may lose DNA if more than 2-3 hours).

The brown protein pellet at the bottom of the tube should dissolve.

If the suspension is still very gelatinous after incubation, add another 100 μ l SSTE and 50 μ l proteinase K.

Separate proteins – Phenol Extraction:

Perform this procedure in the fume hood.

Add an equal volume of phenol and mix by inverting.

Spin for 2 minutes @ 8000 rpm.

Using a pipette, remove the aqueous (upper) phase to another 1.5 ml tube. Leave as much protein (white stringy material) behind as possible, and take none of the lower phase.

Repeat until no protein is visible at the interface and the aqueous phase is colourless.

Repeat using a mixture of phenol:chloroform iso-amyl alcohol (25:24:1).

Repeat using an equal volume of chloroform.

Extract DNA:

If the volume of the aqueous phase after the phenol extraction is over 400 μ l, divide it into more than 1 microfuge tubes.

Estimate the volume liquid and add 0.1 of this amount of 2M NaAc. Add 2 volumes of 100% ice-cold ethanol, and invert the tube several times to mix.

Store on ice at least 15-30 minutes (this solution can be stored on ice indefinitely if so wished). Recover the DNA by centrifugation at 14000 rpm for 10 minutes (at 0°C).

Carefully pour the ethanol into the sink, making sure not to lose the DNA pellet.

Add ~ 700 μ l and recentrifuge at 14000 rpm for 2 minutes (at 4 °C).

Carefully drain the ethanol and store the open tube in a safe spot on the lab bench until dry.

Add 50 μ l TE (less if the pellet is small). Move the TE over the quadrant (using a micropipette tip) where the pellet collected to pick up all DNA that was not in the pellet. Dissolve DNA by flicking tube or incubation at 55 °C.

Store at 4 or -20 °C.

Primer Protocols

Set A Protocol: 10 X Reaction Buffer: 10 μ l
 dNTPs (1.25mM each): 16 μ l
 primers (10pm/ μ l): 3 μ l each
 template (10ng/ μ l): 10 μ l
 MgCl₂ (25 mM/ μ l): 6 μ l
 H₂O: 49.5 μ l
 Taq (1 unit/ μ l): 2.5 μ l
 Oil: 1 drop

Cycling Protocol: Hot Start: 95°C - 5 min 80°C - 5 min 95°C - 2 min (1 cycle)
 Cycling: 95°C - 1 min 68°C - 1 min 72°C - 1 min (35 cycles)
 Extension: 72°C - 10 min
 Storage Soak: 4°C - indefinite

Set B Protocol: 10 X Reaction Buffer: 3 μ l
 MgCl₂ (25 mM/ μ l): 2.4 μ l
 dNTPs (1.25mM each): 4.8 μ l
 primers (10pm/ μ l): 3 μ l each
 template (10ng/ μ l): 1 μ l
 MgCl₂ (25 mM/ μ l): 2.4 μ l
 H₂O: 13.8 μ l
 Taq (1 unit/ μ l): 1 μ l
 Oil: 1 drop

Cycling Protocol: Hot Start: 95°C - 5 min 80°C - 5 min 95°C - 2 min (1 cycle)
 Touch Down Cycling: 95°C - 1 min 65°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 64°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 63°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 62°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 61°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 60°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 59°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 58°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 57°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 56°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 55°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 54°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 53°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 52°C - 1 min 72°C - 1 min (7 cycles)
 Storage Soak: 4°C - indefinite

Set C Protocol:

10 X Reaction Buffer:	3.5 μ l
MgCl ₂ (25 mM/ μ l):	2.8 μ l
dNTPs (1.25mM each):	5.6 μ l
primers (10pm/ μ l):	3 μ l each
template (10ng/ μ l):	1 μ l
H ₂ O:	16.1 μ l
Taq (1 unit/ μ l):	1 μ l
Oil:	1 drop

Cycling Protocol: Hot Start: 95°C - 5 min 80°C - 5 min 95°C - 2 min (1 cycle)

Touch Down Cycling:

95°C - 1 min	65°C - 1 min	72°C - 1 min	(3 cycles)
95°C - 1 min	64°C - 1 min	72°C - 1 min	(3 cycles)
95°C - 1 min	63°C - 1 min	72°C - 1 min	(3 cycles)
95°C - 1 min	62°C - 1 min	72°C - 1 min	(3 cycles)
95°C - 1 min	61°C - 1 min	72°C - 1 min	(3 cycles)
95°C - 1 min	60°C - 1 min	72°C - 1 min	(3 cycles)
95°C - 1 min	59°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	58°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	57°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	56°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	55°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	54°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	53°C - 1 min	72°C - 1 min	(2 cycles)
95°C - 1 min	52°C - 1 min	72°C - 1 min	(5 cycles)

Storage Soak: 4°C - indefinite

Duplex Reaction D&E Protocol:

10 X Reaction Buffer:	3.5 μ l
MgCl ₂ (25 mM/ μ l):	4.6 μ l
dNTPs (1.25mM each):	5.6 μ l
primers (10pm/ μ l):	3 μ l each
template (10ng/ μ l):	1 μ l
H ₂ O:	16.1 μ l
Taq (1 unit/ μ l):	1 μ l
Oil:	1 drop

Cycling Protocol: Hot Start: 95°C - 5 min 80°C - 5 min 95°C - 2 min (1 cycle)
 Touch Down Cycling: 95°C - 1 min 48°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 47°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 46°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 45°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 44°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 43°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 42°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 41°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 40°C - 1 min 72°C - 1 min (28cycles)
 Storage Soak: 4°C - indefinite

Duplex Reaction F&G Protocol: 10 X Reaction Buffer: 3.5 µl
 MgCl₂ (25 mM/µl): 4.3 µl
 dNTPs (1.25mM each): 5.6 µl
 primers (10pm/µl): 3 µl each
 template (10ng/µl): 1 µl
 H₂O: 16.1 µl
 Taq (1 unit/µl): 1 µl
 Oil: 1 drop

Cycling Protocol: Hot Start: 95°C - 5 min 80°C - 5 min 95°C - 2 min (1 cycle)
 Touch Down Cycling: 95°C - 1 min 59°C - 1 min 72°C - 1 min (1 cycles)
 95°C - 1 min 58°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 57°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 56°C - 1 min 72°C - 1 min (3 cycles)
 95°C - 1 min 55°C - 1 min 72°C - 1 min (5 cycles)
 95°C - 1 min 54°C - 1 min 72°C - 1 min (5 cycles)
 95°C - 1 min 53°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 52°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 51°C - 1 min 72°C - 1 min (25 cycles)
 Storage Soak: 4°C - indefinite

Duplex Reaction H&I Protocol:

10 X Reaction Buffer:	3.5 μ l
MgCl ₂ (25 mM/ μ l):	4.3 μ l
dNTPs (1.25mM each):	5.6 μ l
primers (10pm/ μ l):	3 μ l each
template (10ng/ μ l):	1 μ l
H ₂ O:	16.1 μ l
Taq (1 unit/ μ l):	1 μ l
Oil:	1 drop

Cycling Protocol: Hot Start: 95°C - 5 min 80°C - 5 min 95°C - 2 min (1 cycle)
 Touch Down Cycling: 95°C - 1 min 59°C - 1 min 72°C - 1 min (1 cycle)
 95°C - 1 min 58°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 57°C - 1 min 72°C - 1 min (2 cycles)
 95°C - 1 min 56°C - 1 min 72°C - 1 min (3 cycles)
 95°C - 1 min 55°C - 1 min 72°C - 1 min (5 cycles)
 95°C - 1 min 54°C - 1 min 72°C - 1 min (5 cycles)
 95°C - 1 min 53°C - 1 min 72°C - 1 min (5 cycles)
 95°C - 1 min 52°C - 1 min 72°C - 1 min (26 cycles)
 Storage Soak: 4°C - indefinite

Solutions and Buffers:**EDTA**

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

186.12 g $\text{Na}_2\text{EDTA}\cdot\text{H}_2\text{O}$
 800 ml DDW (double distilled H_2O)

M.W. = 372.2 g / mol

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

Proteinase K

Concentration: 20 mg/ml in 50% glycerol
pH: -
Amount: -
Autoclave: No

- Aliquot into autoclaved 1.5 ml epindorf tubes and store at $-20\text{ }^\circ\text{C}$.

Red Cell Lysis Buffer (NH_4Cl Na_4HCO_3)

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

70.0 g NH_4Cl 1.31 M
 0.71 g NH_4HCO_3 0.009M
 800 ml DDW

M.W. = 53.45 g/mol

M.W. = 78.98 g/mol

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

Saline Solution (NaCl)

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

4.5 g NaCl
 400 ml DDW
 - adjust volume to 500 ml with DDW

SDS

Concentration: 10%
pH: -
Amount: 500 ml
Autoclave: No

50 g SDS
 400 ml DDW

M.W. = n/a

- adjust volume to 500 ml with DDW

Sodium Acetate (NaAc)

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

27.22 g NaAc
 80 ml DDW

M.W. = 136.1 g / mol

- adjust volume to 100 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

STE

Concentration: 1 X
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

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

TE

Concentration: 1 X
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

Tris (Tris-HCl)

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

242.3 g Tris
800 ml DDW

M.W. = 121.14

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

Appendix B: India Restriction Data

					Sites				
	663 HaeIII	5176 AluI	7025 AluI	10394 DdeI	10397 AluI	13262 AluI	16065 HinfI	16517 HaeIII	9 bp del
Ind1	0	1	1	1	1	0	1	1	0
Ind2	0	1	1	1	1	0	1	1	0
Ind3	0	1	1	0	0	0	1	1	0
Ind4	0	1	1	1	1	0	1	1	0
Ind5	0	1	1	0	0	0	1	1	0
Ind6	0	1	1	1	1	0	1	1	0
Ind7	0	1	1	0	0	0	1	1	0
Ind8	0	1	1	1	1	0	1	1	0
Ind9	0	1	1	0	0	0	1	1	0
Ind10	0	1	1	1	1	0	1	1	0
Ind11	0	1	1	0	0	0	1	1	0
Ind12	0	1	1	1	1	0	1	1	0
Ind13	0	1	1	1	1	0	1	1	0
Ind14	0	1	1	1	1	0	1	1	0
Ind15	0	1	1	1	1	0	1	1	0
Ind16	0	1	1	1	1	0	1	1	0
Ind17	0	1	1	1	1	0	1	1	0
Ind18	0	1	1	0	0	0	1	1	0
Ind19	0	1	1	0	0	0	1	1	0
Ind20	0	1	1	0	0	0	1	1	0
Ind21	0	1	1	1	1	0	1	0	0
Ind22	0	1	1	1	1	0	1	1	0
Ind23	0	1	1	1	1	0	1	1	0
Ind24	0	1	1	0	0	0	1	1	0
Ind25	0	1	1	1	1	0	1	1	0
Ind26	0	1	1	1	1	0	1	1	0
Ind27	0	1	1	0	0	0	1	0	0
Ind28	0	1	1	0	0	0	1	1	0
Ind29	0	1	1	0	0	0	1	1	0
Ind30	0	1	1	0	0	0	1	1	0
Ind31	0	1	0	0	0	0	1	1	0
Ind32	0	1	1	1	1	0	1	1	0
Ind33	0	1	1	1	1	0	1	1	0

Ind34	0	1	1	0	0	0	1	1	0
Ind35	0	1	1	1	1	0	1	1	0
Ind36	0	1	1	1	1	0	1	1	0
Ind37	0	1	1	0	0	0	1	1	0
Ind38	0	1	1	1	1	0	1	1	0
Ind39	0	1	1	1	1	0	1	1	0
Ind40	0	1	1	1	1	0	1	1	0
Ind41	0	1	0	0	0	0	1	1	0
Ind42	0	1	1	1	1	0	1	1	0
Ind43	0	1	1	1	1	0	1	1	0
Ind44	0	1	1	0	0	0	1	1	0
Ind45	0	1	1	1	1	0	1	1	0
Ind46	0	1	1	1	1	0	1	1	0
Ind47	0	1	1	0	0	0	1	0	0
Ind48	0	1	1	1	1	0	1	1	0
Ind49	0	1	1	1	1	0	1	1	0
Ind50	0	1	1	1	1	0	1	0	0
Ind51	0	1	1	1	1	0	1	1	0
Ind52	1	1	1	0	0	0	1	1	0
Ind53	0	1	1	0	0	0	1	0	0
Ind54	0	1	1	0	0	0	1	0	0
Ind55	0	1	1	0	0	0	1	1	0
Ind56	0	1	1	0	0	0	1	1	0
Ind57	0	1	1	0	0	0	1	1	0
Ind58	0	1	1	0	0	0	1	1	0
Ind59	0	1	1	0	0	0	1	1	0
Ind60	0	1	0	0	0	0	1	1	0
Ind61	0	1	1	0	0	0	1	1	0
Ind62	0	1	1	0	0	0	1	0	0
Ind63	0	1	1	0	0	0	1	1	0
Ind64	0	1	1	1	1	0	1	1	0
Ind65	0	1	1	1	1	0	1	1	0
Ind66	0	1	1	0	0	0	1	1	0
Ind67	0	1	1	1	1	0	1	1	0
Ind68	0	1	1	1	1	0	1	1	0
Ind69	0	1	1	0	0	0	1	1	0
Ind70	0	1	1	0	0	0	1	0	0
Ind71	0	1	1	0	0	0	1	1	0
Ind72	0	1	1	1	1	0	1	1	0
Ind73	0	1	1	1	1	0	1	1	0

Ind74	0	1	1	1	1	0	1	1	0
Ind75	0	1	1	0	0	0	1	1	0
Ind76	0	1	1	1	1	0	1	1	0
Ind77	0	1	1	1	1	0	1	1	0
Ind78	0	1	1	1	1	0	1	1	0
Ind79	0	1	1	1	1	0	1	1	1
Ind80	0	1	1	0	0	0	1	1	0
Ind81	0	1	1	0	0	0	1	1	0
Ind82	0	1	1	1	1	0	1	0	0
Ind83	0	1	1	1	1	0	1	1	0
Ind84	0	1	1	1	1	0	1	1	0
Ind85	0	1	1	0	0	0	1	1	0
Ind86	0	1	1	1	1	0	1	1	0
Ind87	0	1	1	0	0	0	1	0	0
Ind88	0	1	1	0	0	0	1	1	0
Ind89	0	1	1	0	0	0	1	1	0
Ind90	0	0	1	0	0	0	1	1	0
Ind91	0	1	1	1	1	0	1	1	0
Ind92	0	1	1	1	1	0	1	1	0
Ind93	0	1	1	1	1	0	1	1	0
Ind94	0	1	1	1	1	0	1	1	0
Ind95	0	1	1	1	1	0	1	1	0
Ind96	0	1	1	1	1	0	1	1	0
Ind97	0	1	1	1	1	0	1	0	0
Ind98	0	1	1	0	0	0	1	1	0
Ind99	0	1	1	1	1	0	1	1	0
Ind100	0	1	1	1	1	0	1	1	0
Ind101	0	1	1	0	0	0	1	0	0
Ind102	0	1	1	1	1	0	1	0	0
Ind103	0	1	1	0	0	0	1	1	0
Ind104	0	1	1	0	0	1	1	1	0
Ind105	0	1	1	1	1	0	1	1	0
Ind106	0	1	1	1	1	0	1	1	0
Ind107	0	1	1	1	1	0	1	1	0
Ind108	0	1	1	1	1	0	1	1	0
Ind109	0	1	1	1	1	0	1	1	0
Ind110	0	1	1	1	1	0	?	?	0
Ind111	0	1	1	0	0	0	1	0	0
Ind112	0	1	1	0	0	0	1	0	0
Ind113	0	1	1	1	1	0	1	1	0

Ind114	0	1	1	1	1	0	1	0	0
Ind115	0	1	1	0	0	0	1	0	0
Ind116	0	1	1	0	0	0	1	1	0
Ind117	0	1	1	0	0	0	1	1	0
Ind118	0	1	1	1	1	0	1	0	0
Ind119	0	1	1	1	1	0	1	1	0
Ind120	0	1	1	1	1	0	1	1	0
Ind121	0	1	1	1	1	0	1	1	0
Ind122	0	1	1	1	1	0	1	0	0
Ind123	0	1	1	0	0	0	1	1	0
Ind124	0	1	1	1	1	0	1	1	0
Ind125	0	1	1	0	0	0	1	1	0
Ind126	0	1	1	0	0	0	1	0	0
Ind127	0	1	1	1	1	0	1	1	0
Ind128	0	1	1	0	0	0	1	1	0
Ind129	0	1	1	1	1	0	1	1	0
Ind130	0	1	1	1	1	0	1	1	0
Ind131	0	1	1	1	1	0	1	1	0
Ind132	0	1	1	1	1	0	1	1	0
Ind133	0	1	1	1	1	0	1	0	0
Ind134	0	1	1	1	1	0	1	1	0
Ind135	0	1	1	1	1	0	1	1	0
Ind136	0	1	1	0	0	0	1	1	0
Ind137	0	1	1	1	1	0	1	0	0
Ind138	0	1	1	0	0	0	1	1	0
Ind139	0	1	1	0	0	0	1	1	0
Ind140	0	1	1	1	1	0	1	1	0
Ind141	0	1	1	0	0	0	1	1	0
Ind142	0	1	1	1	1	0	1	1	0
Ind143	0	1	1	0	0	0	1	1	0
Ind144	0	1	1	0	0	0	1	1	0
Ind145	0	1	1	1	1	0	1	1	0
Ind146	0	1	1	0	0	0	1	0	0
Ind147	0	1	1	1	1	0	1	1	0
Ind148	0	1	1	1	1	0	1	1	0
Ind149	0	1	1	1	1	0	1	1	0
Ind150	0	1	1	1	1	0	1	1	0
Ind151	0	1	1	0	0	0	1	1	0
Ind152	0	1	1	1	0	0	1	0	0
Ind153	0	1	1	1	1	0	1	1	0

Ind154	0	1	1	1	1	0	1	1	0
Ind155	0	1	1	1	1	0	1	1	0
Ind156	0	1	1	0	0	0	1	1	0
Ind157	0	1	1	0	0	0	1	1	0
Ind158	0	1	0	0	0	0	1	0	0
Ind159	0	1	1	1	1	0	1	1	0
Ind160	0	1	1	0	0	0	1	0	0
Ind161	0	1	1	1	1	0	1	1	0
Ind162	0	1	1	1	1	0	1	1	0
Ind163	0	1	1	0	0	0	1	1	0
Ind164	0	1	1	1	1	0	1	1	0
Ind165	0	1	1	0	0	0	1	0	0
Ind166	0	1	1	1	0	0	1	1	0
Ind168	0	1	1	0	0	0	1	1	0
Ind169	0	1	1	1	1	0	1	1	0
Ind170	0	1	1	0	0	0	1	1	0
Ind171	0	1	1	0	0	0	1	0	0
Ind172	0	1	1	1	1	0	1	1	0
Ind173	0	1	1	0	0	0	1	1	0
Ind174	0	1	0	0	0	0	1	0	0
Ind175	0	1	0	0	0	0	1	1	0
Ind176	0	1	1	0	0	0	1	0	0
Ind177	0	1	1	1	1	0	1	1	0
Ind178	0	1	1	0	0	0	1	1	0
Ind179	0	1	1	0	0	0	1	0	0
Ind180	0	1	1	0	0	0	1	0	0
Ind181	0	1	1	0	0	0	1	1	0
Ind182	0	1	1	1	1	0	1	1	1
Ind183	0	1	0	0	0	0	1	0	0
Ind184	0	1	1	0	0	0	1	1	0
Ind185	0	1	1	0	0	0	1	1	0
Ind186	0	1	1	0	0	0	1	1	0
Ind187	0	1	1	0	0	0	1	0	0
Ind188	0	1	1	0	0	0	1	1	0
Ind189	0	1	1	1	1	0	0	1	0

Appendix C: Sorted Total Data

					Sites						
	663	5176	7025	10394	10397	13262	16065	16517	9 bp		
Sample	HaeIII	AluI	AluI	DdeI	AluI	AluI	HinfI	HaeIII	del	Classification	
Haplotype 1											
TorT134	0	0	0	1	1	1	1	0	0		
Haplotype 2											
Ind90	0	0	1	0	0	0	1	1	0		
Haplotype 3											
TorT131	0	0	1	0	0	1	1	0	0	*2	
TorS16	0	0	1	0	0	1	1	0	0	EVE	
TorNA50	0	0	1	0	0	1	1	0	0		
Haplotype 4											
Bal110	0	0	1	0	0	1	1	1	0	SA	
Bal121	0	0	1	0	0	1	1	1	0	SA	
TorT135	0	0	1	0	0	1	1	1	0		
Haplotype 5											
Bal57	0	0	1	0	0	1	1	1	1	TW,TW	
Haplotype 6											
TorI63	0	0	1	1	1	0	1	0	0		
Haplotype 7											
Bal25	0	0	1	1	1	1	1	0	0	MC*2, TW,KN	
Bal97	0	0	1	1	1	1	1	0	0	KN	
TorT132	0	0	1	1	1	1	1	0	0		
TorS10	0	0	1	1	1	1	1	0	0	NIV*13	
TorS12	0	0	1	1	1	1	1	0	0	NIV*2	
TorS13	0	0	1	1	1	1	1	0	0	EVE*2	
TorS14	0	0	1	1	1	1	1	0	0	EVE	

TorS15	0	0	1	1	1	1	1	0	0	EVE
TorNA46	0	0	1	1	1	1	1	0	0	
TorNA47	0	0	1	1	1	1	1	0	0	
TorNA48	0	0	1	1	1	1	1	0	0	
TorNA49	0	0	1	1	1	1	1	0	0	
TorNA88	0	0	1	1	1	1	1	0	0	
TorNA89	0	0	1	1	1	1	1	0	0	
TorNA90	0	0	1	1	1	1	1	0	0	
TorNA92	0	0	1	1	1	1	1	0	0	
TorNA93	0	0	1	1	1	1	1	0	0	
TorNA95	0	0	1	1	1	1	1	0	0	
Haplotype 8										
Bal102	0	0	1	1	1	1	1	1	0	KN
TorT133	0	0	1	1	1	1	1	1	0	
TorT136	0	0	1	1	1	1	1	1	0	
TorT137	0	0	1	1	1	1	1	1	0	
TorT138	0	0	1	1	1	1	1	1	0	
TorS11	0	0	1	1	1	1	1	1	0	NIV
TorNA44	0	0	1	1	1	1	1	1	0	
TorNA45	0	0	1	1	1	1	1	1	0	
TorNA91	0	0	1	1	1	1	1	1	0	
TorNA94	0	0	1	1	1	1	1	1	0	
TorNA96	0	0	1	1	1	1	1	1	0	
Haplotype 9										
Ind158	0	1	0	0	0	0	1	0	0	
Ind174	0	1	0	0	0	0	1	0	0	
Ind183	0	1	0	0	0	0	1	0	0	
TorE1	0	1	0	0	0	0	1	0	0	FIN*2
TorE6	0	1	0	0	0	0	1	0	0	FIN
TorE7	0	1	0	0	0	0	1	0	0	FIN
TorE9	0	1	0	0	0	0	1	0	0	SWE
TorE10	0	1	0	0	0	0	1	0	0	SWE
TorE11	0	1	0	0	0	0	1	0	0	FIN*2, SWE
TorE13	0	1	0	0	0	0	1	0	0	SWE
TorE14	0	1	0	0	0	0	1	0	0	SWE
TorE15	0	1	0	0	0	0	1	0	0	SWE*2
TorE16	0	1	0	0	0	0	1	0	0	SWE

TorC8	0	1	0	0	0	0	1	0	0	
TorC13	0	1	0	0	0	0	1	0	0	
TorC14	0	1	0	0	0	0	1	0	0	
TorC21	0	1	0	0	0	0	1	0	0	
TorC31	0	1	0	0	0	0	1	0	0	
TorC32	0	1	0	0	0	0	1	0	0	
TorC33	0	1	0	0	0	0	1	0	0	
TorC37	0	1	0	0	0	0	1	0	0	
TorC47	0	1	0	0	0	0	1	0	0	
TorC48	0	1	0	0	0	0	1	0	0	
TorC56	0	1	0	0	0	0	1	0	0	
TorC74	0	1	0	0	0	0	1	0	0	
TorC105	0	1	0	0	0	0	1	0	0	
TorC111	0	1	0	0	0	0	1	0	0	
TorC112	0	1	0	0	0	0	1	0	0	
TorC116	0	1	0	0	0	0	1	0	0	
TorI2	0	1	0	0	0	0	1	0	0	
TorI4	0	1	0	0	0	0	1	0	0	
TorI11	0	1	0	0	0	0	1	0	0	
Haplotype 10										
Ind31	0	1	0	0	0	0	1	1	0	
Ind41	0	1	0	0	0	0	1	1	0	
Ind60	0	1	0	0	0	0	1	1	0	
Ind175	0	1	0	0	0	0	1	1	0	
TorE2	0	1	0	0	0	0	1	1	0	FIN*9, SWE*4
TorE3	0	1	0	0	0	0	1	1	0	SWE*2
TorE4	0	1	0	0	0	0	1	1	0	FIN
TorE5	0	1	0	0	0	0	1	1	0	FIN
TorE8	0	1	0	0	0	0	1	1	0	FIN
TorE12	0	1	0	0	0	0	1	1	0	FIN
TorE17	0	1	0	0	0	0	1	1	0	SWE
TorE18	0	1	0	0	0	0	1	1	0	FIN
TorC1	0	1	0	0	0	0	1	1	0	
TorC2	0	1	0	0	0	0	1	1	0	
TorC3	0	1	0	0	0	0	1	1	0	
TorC4	0	1	0	0	0	0	1	1	0	
TorC6	0	1	0	0	0	0	1	1	0	
TorC11	0	1	0	0	0	0	1	1	0	

TorC17	0	1	0	0	0	0	1	1	0	
TorC18	0	1	0	0	0	0	1	1	0	
TorC19	0	1	0	0	0	0	1	1	0	
TorC20	0	1	0	0	0	0	1	1	0	
TorC22	0	1	0	0	0	0	1	1	0	
TorC24	0	1	0	0	0	0	1	1	0	
TorC27	0	1	0	0	0	0	1	1	0	
TorC30	0	1	0	0	0	0	1	1	0	
TorC34	0	1	0	0	0	0	1	1	0	
TorC35	0	1	0	0	0	0	1	1	0	
TorC38	0	1	0	0	0	0	1	1	0	
TorC39	0	1	0	0	0	0	1	1	0	
TorC40	0	1	0	0	0	0	1	1	0	
TorC44	0	1	0	0	0	0	1	1	0	
TorC50	0	1	0	0	0	0	1	1	0	
TorC52	0	1	0	0	0	0	1	1	0	
TorC54	0	1	0	0	0	0	1	1	0	
TorC55	0	1	0	0	0	0	1	1	0	
TorC100	0	1	0	0	0	0	1	1	0	
TorC113	0	1	0	0	0	0	1	1	0	
TorC114	0	1	0	0	0	0	1	1	0	
TorC115	0	1	0	0	0	0	1	1	0	
TorI1	0	1	0	0	0	0	1	1	0	*20
TorI3	0	1	0	0	0	0	1	1	0	
TorI5	0	1	0	0	0	0	1	1	0	
TorI6	0	1	0	0	0	0	1	1	0	
TorI7	0	1	0	0	0	0	1	1	0	
TorI8	0	1	0	0	0	0	1	1	0	*2
TorI9	0	1	0	0	0	0	1	1	0	*2
TorI10	0	1	0	0	0	0	1	1	0	
TorI12	0	1	0	0	0	0	1	1	0	
Haplotype 11										
TorNA28	0	1	0	0	0	1	1	1	0	
Haplotype 12										
TorC76	0	1	0	1	0	0	1	0	0	
Haplotype 13										
TorI53	0	1	1	0	0	0	0	1	0	

Haplotype 14										
Ind27	0	1	1	0	0	0	1	0	0	
Ind47	0	1	1	0	0	0	1	0	0	
Ind53	0	1	1	0	0	0	1	0	0	
Ind54	0	1	1	0	0	0	1	0	0	
Ind62	0	1	1	0	0	0	1	0	0	
Ind70	0	1	1	0	0	0	1	0	0	
Ind87	0	1	1	0	0	0	1	0	0	
Ind101	0	1	1	0	0	0	1	0	0	
Ind111	0	1	1	0	0	0	1	0	0	
Ind112	0	1	1	0	0	0	1	0	0	
Ind115	0	1	1	0	0	0	1	0	0	
Ind126	0	1	1	0	0	0	1	0	0	
Ind146	0	1	1	0	0	0	1	0	0	
Ind160	0	1	1	0	0	0	1	0	0	
Ind165	0	1	1	0	0	0	1	0	0	
Ind171	0	1	1	0	0	0	1	0	0	
Ind176	0	1	1	0	0	0	1	0	0	
Ind179	0	1	1	0	0	0	1	0	0	
Ind180	0	1	1	0	0	0	1	0	0	
Ind187	0	1	1	0	0	0	1	0	0	
Chen1	0	1	1	0	0	0	1	0	0	MAN
Chen2	0	1	1	0	0	0	1	0	0	OTH*3
TorE24	0	1	1	0	0	0	1	0	0	FIN,SWE
TorE25	0	1	1	0	0	0	1	0	0	SWE
TorE26	0	1	1	0	0	0	1	0	0	FIN
TorE27	0	1	1	0	0	0	1	0	0	FIN
TorE28	0	1	1	0	0	0	1	0	0	FIN*2
TorE29	0	1	1	0	0	0	1	0	0	FIN
TorE30	0	1	1	0	0	0	1	0	0	FIN
TorE33	0	1	1	0	0	0	1	0	0	SWE
TorE36	0	1	1	0	0	0	1	0	0	SWE
TorE37	0	1	1	0	0	0	1	0	0	SWE
TorE38	0	1	1	0	0	0	1	0	0	FIN
TorE39	0	1	1	0	0	0	1	0	0	FIN
TorE40	0	1	1	0	0	0	1	0	0	FIN
TorC5	0	1	1	0	0	0	1	0	0	
TorC12	0	1	1	0	0	0	1	0	0	
TorC28	0	1	1	0	0	0	1	0	0	

TorC29	0	1	1	0	0	0	1	0	0	
TorC36	0	1	1	0	0	0	1	0	0	
TorC42	0	1	1	0	0	0	1	0	0	
TorC45	0	1	1	0	0	0	1	0	0	
TorC49	0	1	1	0	0	0	1	0	0	
TorC51	0	1	1	0	0	0	1	0	0	
TorC61	0	1	1	0	0	0	1	0	0	
TorC62	0	1	1	0	0	0	1	0	0	
TorC63	0	1	1	0	0	0	1	0	0	
TorC65	0	1	1	0	0	0	1	0	0	
TorC66	0	1	1	0	0	0	1	0	0	
TorC69	0	1	1	0	0	0	1	0	0	
TorC70	0	1	1	0	0	0	1	0	0	
TorC72	0	1	1	0	0	0	1	0	0	
TorC73	0	1	1	0	0	0	1	0	0	
TorC99	0	1	1	0	0	0	1	0	0	
TorI21	0	1	1	0	0	0	1	0	0	*2
TorI22	0	1	1	0	0	0	1	0	0	
TorI23	0	1	1	0	0	0	1	0	0	
TorI25	0	1	1	0	0	0	1	0	0	
TorI27	0	1	1	0	0	0	1	0	0	
TorI28	0	1	1	0	0	0	1	0	0	
TorI30	0	1	1	0	0	0	1	0	0	
TorI32	0	1	1	0	0	0	1	0	0	
TorI33	0	1	1	0	0	0	1	0	0	
TorI34	0	1	1	0	0	0	1	0	0	
TorI35	0	1	1	0	0	0	1	0	0	
TorI37	0	1	1	0	0	0	1	0	0	
TorI40	0	1	1	0	0	0	1	0	0	*4
TorI41	0	1	1	0	0	0	1	0	0	
TorI42	0	1	1	0	0	0	1	0	0	
TorI44	0	1	1	0	0	0	1	0	0	
TorI45	0	1	1	0	0	0	1	0	0	
TorC58	0	1	1	0	0	0	1	0	0	
Haplotype 15										
Ind3	0	1	1	0	0	0	1	1	0	
Ind5	0	1	1	0	0	0	1	1	0	
Ind7	0	1	1	0	0	0	1	1	0	
Ind9	0	1	1	0	0	0	1	1	0	

Ind11	0	1	1	0	0	0	1	1	0
Ind18	0	1	1	0	0	0	1	1	0
Ind19	0	1	1	0	0	0	1	1	0
Ind20	0	1	1	0	0	0	1	1	0
Ind24	0	1	1	0	0	0	1	1	0
Ind28	0	1	1	0	0	0	1	1	0
Ind29	0	1	1	0	0	0	1	1	0
Ind30	0	1	1	0	0	0	1	1	0
Ind34	0	1	1	0	0	0	1	1	0
Ind37	0	1	1	0	0	0	1	1	0
Ind44	0	1	1	0	0	0	1	1	0
Ind55	0	1	1	0	0	0	1	1	0
Ind56	0	1	1	0	0	0	1	1	0
Ind57	0	1	1	0	0	0	1	1	0
Ind58	0	1	1	0	0	0	1	1	0
Ind59	0	1	1	0	0	0	1	1	0
Ind61	0	1	1	0	0	0	1	1	0
Ind63	0	1	1	0	0	0	1	1	0
Ind66	0	1	1	0	0	0	1	1	0
Ind69	0	1	1	0	0	0	1	1	0
Ind71	0	1	1	0	0	0	1	1	0
Ind75	0	1	1	0	0	0	1	1	0
Ind80	0	1	1	0	0	0	1	1	0
Ind81	0	1	1	0	0	0	1	1	0
Ind85	0	1	1	0	0	0	1	1	0
Ind88	0	1	1	0	0	0	1	1	0
Ind89	0	1	1	0	0	0	1	1	0
Ind98	0	1	1	0	0	0	1	1	0
Ind103	0	1	1	0	0	0	1	1	0
Ind116	0	1	1	0	0	0	1	1	0
Ind117	0	1	1	0	0	0	1	1	0
Ind123	0	1	1	0	0	0	1	1	0
Ind125	0	1	1	0	0	0	1	1	0
Ind128	0	1	1	0	0	0	1	1	0
Ind136	0	1	1	0	0	0	1	1	0
Ind138	0	1	1	0	0	0	1	1	0
Ind139	0	1	1	0	0	0	1	1	0
Ind141	0	1	1	0	0	0	1	1	0
Ind143	0	1	1	0	0	0	1	1	0
Ind144	0	1	1	0	0	0	1	1	0

Ind151	0	1	1	0	0	0	1	1	0	
Ind156	0	1	1	0	0	0	1	1	0	
Ind157	0	1	1	0	0	0	1	1	0	
Ind163	0	1	1	0	0	0	1	1	0	
Ind168	0	1	1	0	0	0	1	1	0	
Ind170	0	1	1	0	0	0	1	1	0	
Ind173	0	1	1	0	0	0	1	1	0	
Ind178	0	1	1	0	0	0	1	1	0	
Ind181	0	1	1	0	0	0	1	1	0	
Ind184	0	1	1	0	0	0	1	1	0	
Ind185	0	1	1	0	0	0	1	1	0	
Ind186	0	1	1	0	0	0	1	1	0	
Ind188	0	1	1	0	0	0	1	1	0	
Chen3	0	1	1	0	0	0	1	1	0	PUL
Chen62	0	1	1	0	0	0	1	1	0	WPYG*4
Chen63	0	1	1	0	0	0	1	1	0	WPYG
TorE19	0	1	1	0	0	0	1	1	0	FIN*2, SWE*4
TorE20	0	1	1	0	0	0	1	1	0	FIN
TorE21	0	1	1	0	0	0	1	1	0	SWE
TorE22	0	1	1	0	0	0	1	1	0	SWE
TorE23	0	1	1	0	0	0	1	1	0	SWE*2
TorE31	0	1	1	0	0	0	1	1	0	SWE
TorE32	0	1	1	0	0	0	1	1	0	FIN
TorE34	0	1	1	0	0	0	1	1	0	SWE
TorE35	0	1	1	0	0	0	1	1	0	SWE
TorE41	0	1	1	0	0	0	1	1	0	FIN*2
TorE42	0	1	1	0	0	0	1	1	0	FIN
TorE43	0	1	1	0	0	0	1	1	0	FIN
TorC7	0	1	1	0	0	0	1	1	0	
TorC9	0	1	1	0	0	0	1	1	0	
TorC10	0	1	1	0	0	0	1	1	0	
TorC15	0	1	1	0	0	0	1	1	0	
TorC16	0	1	1	0	0	0	1	1	0	
TorC23	0	1	1	0	0	0	1	1	0	
TorC25	0	1	1	0	0	0	1	1	0	
TorC26	0	1	1	0	0	0	1	1	0	
TorC41	0	1	1	0	0	0	1	1	0	
TorC43	0	1	1	0	0	0	1	1	0	
TorC46	0	1	1	0	0	0	1	1	0	

TorC53	0	1	1	0	0	0	1	1	0	
TorC57	0	1	1	0	0	0	1	1	0	
TorC59	0	1	1	0	0	0	1	1	0	
TorC60	0	1	1	0	0	0	1	1	0	
TorC64	0	1	1	0	0	0	1	1	0	
TorC67	0	1	1	0	0	0	1	1	0	
TorC68	0	1	1	0	0	0	1	1	0	
TorC71	0	1	1	0	0	0	1	1	0	
TorC75	0	1	1	0	0	0	1	1	0	
TorC98	0	1	1	0	0	0	1	1	0	
TorC101	0	1	1	0	0	0	1	1	0	
TorC102	0	1	1	0	0	0	1	1	0	
TorC108	0	1	1	0	0	0	1	1	0	
TorC118	0	1	1	0	0	0	1	1	0	
TorNA30	0	1	1	0	0	0	1	1	0	
TorNA31	0	1	1	0	0	0	1	1	0	
TorI13	0	1	1	0	0	0	1	1	0	*2
TorI14	0	1	1	0	0	0	1	1	0	
TorI15	0	1	1	0	0	0	1	1	0	
TorI16	0	1	1	0	0	0	1	1	0	
TorI17	0	1	1	0	0	0	1	1	0	
TorI18	0	1	1	0	0	0	1	1	0	
TorI19	0	1	1	0	0	0	1	1	0	
TorI20	0	1	1	0	0	0	1	1	0	
TorI24	0	1	1	0	0	0	1	1	0	*4
TorI26	0	1	1	0	0	0	1	1	0	
TorI29	0	1	1	0	0	0	1	1	0	*2
TorI31	0	1	1	0	0	0	1	1	0	
TorI36	0	1	1	0	0	0	1	1	0	
TorI38	0	1	1	0	0	0	1	1	0	
TorI39	0	1	1	0	0	0	1	1	0	
TorI43	0	1	1	0	0	0	1	1	0	
TorI46	0	1	1	0	0	0	1	1	0	
TorI47	0	1	1	0	0	0	1	1	0	
TorI48	0	1	1	0	0	0	1	1	0	
TorI55	0	1	1	0	0	0	1	1	0	
TorI56	0	1	1	0	0	0	1	1	0	
Haplotype 16										
Bal17	0	1	1	0	0	1	1	0	0	MC

Bal18	0	1	1	0	0	1	1	0	0	MC
Bal19	0	1	1	0	0	1	1	0	0	MC
Bal20	0	1	1	0	0	1	1	0	0	MC
Bal24	0	1	1	0	0	1	1	0	0	MC
Bal27	0	1	1	0	0	1	1	0	0	MC
Bal44	0	1	1	0	0	1	1	0	0	VN
Bal67	0	1	1	0	0	1	1	0	0	TW
Bal68	0	1	1	0	0	1	1	0	0	TW
Bal90	0	1	1	0	0	1	1	0	0	MM
Bal112	0	1	1	0	0	1	1	0	0	SA*2
TorT142	0	1	1	0	0	1	1	0	0	
TorT147	0	1	1	0	0	1	1	0	0	
TorS17	0	1	1	0	0	1	1	0	0	NIV
Haplotype 17										
Bal61	0	1	1	0	0	1	1	0	1	TW
Haplotype 18										
Ind104	0	1	1	0	0	1	1	1	0	
Bal23	0	1	1	0	0	1	1	1	0	MC
Bal29	0	1	1	0	0	1	1	1	0	MC
Bal30	0	1	1	0	0	1	1	1	0	VN
Bal32	0	1	1	0	0	1	1	1	0	VN
Bal33	0	1	1	0	0	1	1	1	0	VN, MA*7
Bal35	0	1	1	0	0	1	1	1	0	VN
Bal39	0	1	1	0	0	1	1	1	0	VN
Bal45	0	1	1	0	0	1	1	1	0	VN
Bal46	0	1	1	0	0	1	1	1	0	VN
Bal47	0	1	1	0	0	1	1	1	0	VN
Bal48	0	1	1	0	0	1	1	1	0	VN
Bal50	0	1	1	0	0	1	1	1	0	VN
Bal51	0	1	1	0	0	1	1	1	0	VN,TW
Bal53	0	1	1	0	0	1	1	1	0	VN
Bal66	0	1	1	0	0	1	1	1	0	TW
Bal69	0	1	1	0	0	1	1	1	0	MA
Bal71	0	1	1	0	0	1	1	1	0	MA
Bal72	0	1	1	0	0	1	1	1	0	MA*4
Bal74	0	1	1	0	0	1	1	1	0	MA*2
Bal80	0	1	1	0	0	1	1	1	0	MA

Bal85	0	1	1	0	0	1	1	1	0	MM
Bal86	0	1	1	0	0	1	1	1	0	MM
Bal91	0	1	1	0	0	1	1	1	0	MM
Bal93	0	1	1	0	0	1	1	1	0	MM
Bal98	0	1	1	0	0	1	1	1	0	KN
Bal99	0	1	1	0	0	1	1	1	0	KN
Bal111	0	1	1	0	0	1	1	1	0	SA*3
Bal120	0	1	1	0	0	1	1	1	0	SA
TorT118	0	1	1	0	0	1	1	1	0	*2
TorT143	0	1	1	0	0	1	1	1	0	
TorT144	0	1	1	0	0	1	1	1	0	
TorT145	0	1	1	0	0	1	1	1	0	
TorT146	0	1	1	0	0	1	1	1	0	
TorT154	0	1	1	0	0	1	1	1	0	
TorS20	0	1	1	0	0	1	1	1	0	EVE
TorS24	0	1	1	0	0	1	1	1	0	UDE*13
TorNA29	0	1	1	0	0	1	1	1	0	
TorNA74	0	1	1	0	0	1	1	1	0	
TorNA75	0	1	1	0	0	1	1	1	0	
TorNA76	0	1	1	0	0	1	1	1	0	
Haplotype 19										
Bal49	0	1	1	0	0	1	1	1	1	VN,MM
Bal54	0	1	1	0	0	1	1	1	1	VN, TW,SA
Bal58	0	1	1	0	0	1	1	1	1	TW
Bal59	0	1	1	0	0	1	1	1	1	TW
Bal79	0	1	1	0	0	1	1	1	1	MA
Bal84	0	1	1	0	0	1	1	1	1	MM
Bal101	0	1	1	0	0	1	1	1	1	KN
Bal115	0	1	1	0	0	1	1	1	1	SA*3
Bal116	0	1	1	0	0	1	1	1	1	SA
TorT128	0	1	1	0	0	1	1	1	1	
TorT129	0	1	1	0	0	1	1	1	1	
TorT130	0	1	1	0	0	1	1	1	1	
TorNA13	0	1	1	0	0	1	1	1	1	
TorNA14	0	1	1	0	0	1	1	1	1	
TorNA15	0	1	1	0	0	1	1	1	1	
TorNA16	0	1	1	0	0	1	1	1	1	
TorNA17	0	1	1	0	0	1	1	1	1	

TorNA18	0	1	1	0	0	1	1	1	1	
TorNA19	0	1	1	0	0	1	1	1	1	
TorNA20	0	1	1	0	0	1	1	1	1	
TorNA21	0	1	1	0	0	1	1	1	1	
TorNA22	0	1	1	0	0	1	1	1	1	
TorNA23	0	1	1	0	0	1	1	1	1	
TorNA24	0	1	1	0	0	1	1	1	1	
TorNA25	0	1	1	0	0	1	1	1	1	
TorNA26	0	1	1	0	0	1	1	1	1	
TorNA27	0	1	1	0	0	1	1	1	1	
TorNA65	0	1	1	0	0	1	1	1	1	
TorNA66	0	1	1	0	0	1	1	1	1	
TorNA67	0	1	1	0	0	1	1	1	1	
TorNA68	0	1	1	0	0	1	1	1	1	
TorNA69	0	1	1	0	0	1	1	1	1	
TorNA70	0	1	1	0	0	1	1	1	1	
Haplotype 20										
TorE47	0	1	1	1	0	0	0	0	0	FIN*2
TorE48	0	1	1	1	0	0	0	0	0	SWE
TorE49	0	1	1	1	0	0	0	0	0	FIN
TorE51	0	1	1	1	0	0	0	0	0	FIN
TorC79	0	1	1	1	0	0	0	0	0	
TorC81	0	1	1	1	0	0	0	0	0	
TorC89	0	1	1	1	0	0	0	0	0	
TorC97	0	1	1	1	0	0	0	0	0	
TorC106	0	1	1	1	0	0	0	0	0	
TorC107	0	1	1	1	0	0	0	0	0	
TorI58	0	1	1	1	0	0	0	0	0	
TorI59	0	1	1	1	0	0	0	0	0	
TorI60	0	1	1	1	0	0	0	0	0	
TorI61	0	1	1	1	0	0	0	0	0	*2
Haplotype 21										
TorE46	0	1	1	1	0	0	0	1	0	FIN*2
TorE50	0	1	1	1	0	0	0	1	0	FIN
TorC78	0	1	1	1	0	0	0	1	0	
TorC84	0	1	1	1	0	0	0	1	0	
TorC103	0	1	1	1	0	0	0	1	0	
TorI57	0	1	1	1	0	0	0	1	0	

Tor162	0	1	1	1	0	0	0	1	0	
Haplotype 22										
Ind152	0	1	1	1	0	0	1	0	0	
Chen4	0	1	1	1	0	0	1	0	0	WOL, OTH
Chen8	0	1	1	1	0	0	1	0	0	WOL
Chen9	0	1	1	1	0	0	1	0	0	MAN
Chen10	0	1	1	1	0	0	1	0	0	PUL
Chen12	0	1	1	1	0	0	1	0	0	MAN
Chen25	0	1	1	1	0	0	1	0	0	EPYG*2
Chen26	0	1	1	1	0	0	1	0	0	MAN
Chen27	0	1	1	1	0	0	1	0	0	WOL
Chen28	0	1	1	1	0	0	1	0	0	WOL
Chen29	0	1	1	1	0	0	1	0	0	MAN
Chen30	0	1	1	1	0	0	1	0	0	MAN
Chen31	0	1	1	1	0	0	1	0	0	MAN
Chen46	0	1	1	1	0	0	1	0	0	WOL*2, PUL*2
Chen47	0	1	1	1	0	0	1	0	0	WOL
Chen48	0	1	1	1	0	0	1	0	0	MAN
Chen49	0	1	1	1	0	0	1	0	0	MAN*11
Chen50	0	1	1	1	0	0	1	0	0	MAN
Chen51	0	1	1	1	0	0	1	0	0	OTH
Chen52	0	1	1	1	0	0	1	0	0	WOL
Chen53	0	1	1	1	0	0	1	0	0	WOL
Chen54	0	1	1	1	0	0	1	0	0	MAN*2
Chen55	0	1	1	1	0	0	1	0	0	MAN
Chen67	0	1	1	1	0	0	1	0	0	WPYG
Chen68	0	1	1	1	0	0	1	0	0	WPYG
TorC77	0	1	1	1	0	0	1	0	0	
TorC91	0	1	1	1	0	0	1	0	0	
Haplotype 23										
Ind166	0	1	1	1	0	0	1	1	0	
Chen5	0	1	1	1	0	0	1	1	0	MAN*3
Chen6	0	1	1	1	0	0	1	1	0	MAN
Chen7	0	1	1	1	0	0	1	1	0	OTH
Chen11	0	1	1	1	0	0	1	1	0	MAN*2
Chen13	0	1	1	1	0	0	1	1	0	WOL

Chen14	0	1	1	1	0	0	1	1	0	MAN*2, OTH
Chen15	0	1	1	1	0	0	1	1	0	PUL
Chen16	0	1	1	1	0	0	1	1	0	WOL
Chen17	0	1	1	1	0	0	1	1	0	WOL
Chen18	0	1	1	1	0	0	1	1	0	MAN
Chen19	0	1	1	1	0	0	1	1	0	WOL
Chen20	0	1	1	1	0	0	1	1	0	MAN
Chen21	0	1	1	1	0	0	1	1	0	OTH
Chen22	0	1	1	1	0	0	1	1	0	MAN*2
Chen23	0	1	1	1	0	0	1	1	0	MAN
Chen32	0	1	1	1	0	0	1	1	0	MAN*2, WOL*2
Chen33	0	1	1	1	0	0	1	1	0	PUL
Chen34	0	1	1	1	0	0	1	1	0	MAN
Chen35	0	1	1	1	0	0	1	1	0	MAN, WOL
Chen36	0	1	1	1	0	0	1	1	0	MAN
Chen37	0	1	1	1	0	0	1	1	0	EPYG*4
Chen38	0	1	1	1	0	0	1	1	0	EPYG*6
Chen39	0	1	1	1	0	0	1	1	0	EPYG
Chen40	0	1	1	1	0	0	1	1	0	EPYG
Chen41	0	1	1	1	0	0	1	1	0	EPYG
Chen42	0	1	1	1	0	0	1	1	0	EPYG
Chen43	0	1	1	1	0	0	1	1	0	WPYG
Chen44	0	1	1	1	0	0	1	1	0	WPYG
Chen45	0	1	1	1	0	0	1	1	0	OTH
Chen56	0	1	1	1	0	0	1	1	0	PUL
Chen57	0	1	1	1	0	0	1	1	0	OTH
Chen58	0	1	1	1	0	0	1	1	0	MAN*2
Chen59	0	1	1	1	0	0	1	1	0	MAN
Chen64	0	1	1	1	0	0	1	1	0	OTH
Chen65	0	1	1	1	0	0	1	1	0	MAN*3
Chen66	0	1	1	1	0	0	1	1	0	WPYG*3
Chen69	0	1	1	1	0	0	1	1	0	WPYG
Chen70	0	1	1	1	0	0	1	1	0	WOL
Chen71	0	1	1	1	0	0	1	1	0	MAN*8, WOL, PUL
Chen72	0	1	1	1	0	0	1	1	0	WOL

Haplotype 26										
Bal34	0	1	1	1	0	1	1	1	0	VN
Bal105	0	1	1	1	0	1	1	1	0	KN
TorS1	0	1	1	1	0	1	1	1	0	NIV*26, UDE*2
TorS2	0	1	1	1	0	1	1	1	0	NIV
TorS3	0	1	1	1	0	1	1	1	0	NIV*4
TorS4	0	1	1	1	0	1	1	1	0	NIV*5
TorS5	0	1	1	1	0	1	1	1	0	NIV
TorS6	0	1	1	1	0	1	1	1	0	UDE
TorS7	0	1	1	1	0	1	1	1	0	UDE
Haplotype 27										
Bal36	0	1	1	1	0	1	1	1	1	VN
Bal55	0	1	1	1	0	1	1	1	1	VN,TW, SA
Bal60	0	1	1	1	0	1	1	1	1	TW
Bal100	0	1	1	1	0	1	1	1	1	KN
Haplotype 28										
Ind189	0	1	1	1	1	0	0	1	0	
Haplotype 29										
Ind21	0	1	1	1	1	0	1	0	0	
Ind50	0	1	1	1	1	0	1	0	0	
Ind82	0	1	1	1	1	0	1	0	0	
Ind97	0	1	1	1	1	0	1	0	0	
Ind102	0	1	1	1	1	0	1	0	0	
Ind114	0	1	1	1	1	0	1	0	0	
Ind118	0	1	1	1	1	0	1	0	0	
Ind122	0	1	1	1	1	0	1	0	0	
Ind133	0	1	1	1	1	0	1	0	0	
Ind137	0	1	1	1	1	0	1	0	0	
TorNA32	0	1	1	1	1	0	1	0	0	
TorNA33	0	1	1	1	1	0	1	0	0	
TorNA34	0	1	1	1	1	0	1	0	0	
TorNA35	0	1	1	1	1	0	1	0	0	
TorNA36	0	1	1	1	1	0	1	0	0	
TorNA37	0	1	1	1	1	0	1	0	0	
TorNA38	0	1	1	1	1	0	1	0	0	

TorNA40	0	1	1	1	1	0	1	0	0
TorNA41	0	1	1	1	1	0	1	0	0
TorNA77	0	1	1	1	1	0	1	0	0
TorNA78	0	1	1	1	1	0	1	0	0
TorNA81	0	1	1	1	1	0	1	0	0
TorNA82	0	1	1	1	1	0	1	0	0
TorNA84	0	1	1	1	1	0	1	0	0
TorNA85	0	1	1	1	1	0	1	0	0
TorNA87	0	1	1	1	1	0	1	0	0
Haplotype 30									
Ind1	0	1	1	1	1	0	1	1	0
Ind2	0	1	1	1	1	0	1	1	0
Ind4	0	1	1	1	1	0	1	1	0
Ind6	0	1	1	1	1	0	1	1	0
Ind8	0	1	1	1	1	0	1	1	0
Ind10	0	1	1	1	1	0	1	1	0
Ind12	0	1	1	1	1	0	1	1	0
Ind13	0	1	1	1	1	0	1	1	0
Ind14	0	1	1	1	1	0	1	1	0
Ind15	0	1	1	1	1	0	1	1	0
Ind16	0	1	1	1	1	0	1	1	0
Ind17	0	1	1	1	1	0	1	1	0
Ind22	0	1	1	1	1	0	1	1	0
Ind23	0	1	1	1	1	0	1	1	0
Ind25	0	1	1	1	1	0	1	1	0
Ind26	0	1	1	1	1	0	1	1	0
Ind32	0	1	1	1	1	0	1	1	0
Ind33	0	1	1	1	1	0	1	1	0
Ind35	0	1	1	1	1	0	1	1	0
Ind36	0	1	1	1	1	0	1	1	0
Ind38	0	1	1	1	1	0	1	1	0
Ind39	0	1	1	1	1	0	1	1	0
Ind40	0	1	1	1	1	0	1	1	0
Ind42	0	1	1	1	1	0	1	1	0
Ind43	0	1	1	1	1	0	1	1	0
Ind45	0	1	1	1	1	0	1	1	0
Ind46	0	1	1	1	1	0	1	1	0
Ind48	0	1	1	1	1	0	1	1	0
Ind49	0	1	1	1	1	0	1	1	0

Ind51	0	1	1	1	1	0	1	1	0	
Ind64	0	1	1	1	1	0	1	1	0	
Ind65	0	1	1	1	1	0	1	1	0	
Ind67	0	1	1	1	1	0	1	1	0	
Ind68	0	1	1	1	1	0	1	1	0	
Ind72	0	1	1	1	1	0	1	1	0	
Ind73	0	1	1	1	1	0	1	1	0	
Ind74	0	1	1	1	1	0	1	1	0	
Ind76	0	1	1	1	1	0	1	1	0	
Ind77	0	1	1	1	1	0	1	1	0	
Ind78	0	1	1	1	1	0	1	1	0	
Ind83	0	1	1	1	1	0	1	1	0	
Ind84	0	1	1	1	1	0	1	1	0	
Ind86	0	1	1	1	1	0	1	1	0	
Ind91	0	1	1	1	1	0	1	1	0	
Ind92	0	1	1	1	1	0	1	1	0	
Ind93	0	1	1	1	1	0	1	1	0	
Ind94	0	1	1	1	1	0	1	1	0	
Ind95	0	1	1	1	1	0	1	1	0	
Ind96	0	1	1	1	1	0	1	1	0	
Ind99	0	1	1	1	1	0	1	1	0	
Ind100	0	1	1	1	1	0	1	1	0	
Ind105	0	1	1	1	1	0	1	1	0	
Ind106	0	1	1	1	1	0	1	1	0	
Ind107	0	1	1	1	1	0	1	1	0	
Ind108	0	1	1	1	1	0	1	1	0	
Ind109	0	1	1	1	1	0	1	1	0	
Ind113	0	1	1	1	1	0	1	1	0	
Ind119	0	1	1	1	1	0	1	1	0	
Ind120	0	1	1	1	1	0	1	1	0	
Ind121	0	1	1	1	1	0	1	1	0	
Ind124	0	1	1	1	1	0	1	1	0	
Ind127	0	1	1	1	1	0	1	1	0	
Ind129	0	1	1	1	1	0	1	1	0	
Ind130	0	1	1	1	1	0	1	1	0	
Ind131	0	1	1	1	1	0	1	1	0	
Ind132	0	1	1	1	1	0	1	1	0	
Ind134	0	1	1	1	1	0	1	1	0	
Ind135	0	1	1	1	1	0	1	1	0	
Ind140	0	1	1	1	1	0	1	1	0	

Ind142	0	1	1	1	1	0	1	1	0	
Ind145	0	1	1	1	1	0	1	1	0	
Ind147	0	1	1	1	1	0	1	1	0	
Ind148	0	1	1	1	1	0	1	1	0	
Ind149	0	1	1	1	1	0	1	1	0	
Ind150	0	1	1	1	1	0	1	1	0	
Ind153	0	1	1	1	1	0	1	1	0	
Ind154	0	1	1	1	1	0	1	1	0	
Ind155	0	1	1	1	1	0	1	1	0	
Ind159	0	1	1	1	1	0	1	1	0	
Ind161	0	1	1	1	1	0	1	1	0	
Ind162	0	1	1	1	1	0	1	1	0	
Ind164	0	1	1	1	1	0	1	1	0	
Ind169	0	1	1	1	1	0	1	1	0	
Ind172	0	1	1	1	1	0	1	1	0	
Ind177	0	1	1	1	1	0	1	1	0	
Bal65	0	1	1	1	1	0	1	1	0	TW
Chen24	0	1	1	1	1	0	1	1	0	OTH
TorE52	0	1	1	1	1	0	1	1	0	FIN
TorT65	0	1	1	1	1	0	1	1	0	*2
TorS26	0	1	1	1	1	0	1	1	0	EVE*11
TorS27	0	1	1	1	1	0	1	1	0	EVE*6, UDE*7
TorS28	0	1	1	1	1	0	1	1	0	EVE*7
TorS29	0	1	1	1	1	0	1	1	0	EVE*10
TorS30	0	1	1	1	1	0	1	1	0	EVE*2, UDE
TorS31	0	1	1	1	1	0	1	1	0	EVE*2
TorS32	0	1	1	1	1	0	1	1	0	EVE*2
TorS33	0	1	1	1	1	0	1	1	0	EVE*2
TorS34	0	1	1	1	1	0	1	1	0	EVE
TorNA39	0	1	1	1	1	0	1	1	0	
TorNA42	0	1	1	1	1	0	1	1	0	
TorNA43	0	1	1	1	1	0	1	1	0	
TorNA79	0	1	1	1	1	0	1	1	0	
TorNA80	0	1	1	1	1	0	1	1	0	
TorNA86	0	1	1	1	1	0	1	1	0	
Haplotype 31										
Ind79	0	1	1	1	1	0	1	1	1	

Haplotype 41									
TorNA7	1	1	1	1	0	1	1	1	0
TorNA8	1	1	1	1	0	1	1	1	0
Literature Sources:				Classifications:				9 bp deletion key:	
Bal - Ballinger et al, 1992				MC = Malaysian Chinese				1= 9 bp deletion	
Chen - Chen et al, 1995				VA = Vietnamese				present	
TorC - Torroni et al, 1994a				MA = Malay Aborigines				0= 9 bp deletion	
TorE - Torroni et al, 1996				MM =Malays				absent	
TorI - Torroni et al, 1997				TW = TaiwaneseHan					
TorNA - Torroni et al, 1993a				KN =Koreans					
TorS - Torroni et al, 1993b				SA = Sabah Aborigines					
TorT - Torroni et al, 1994b				MAN = Mandenkalu					
				WOL =Wolof					
				PUL = Pular					
				OTH = Other Senegalese					
				EPYG = Eastern Pygmies					
				WPGY = Western Pygmies					
				FIN = Finn					
				SWE =Swede					
				NIV = Nivkhs					
				EVE =Evenks					
				UDE = Udegeys					

References

- Agrawal, D.P. (1996) *The Vedic Hapappans*. (review)
<http://www.picatype.com/dig/dm2/dm2aa07.htm>
- Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R. and Young, I.G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290(5806)**: 457-465.
- Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A. and Saunders, N.C. (1987) Intraspecific Phylogeography: The Mitochondrial DNA Bridge Between Population Genetics and Systematics. *Annual Review of Ecology and Systematics* **18**: 489-522.
- Balakrishnan, V. (1978) A Preliminary Study of Genetic Distances among Some Populations of the Indian Sub-continent. *Journal of Human Evolution* **7**: 67-75.
- Ballinger, S.W., Schurr, T.G., Torroni, A., Gan, Y.Y., Hodge, J.A., Hassan, K., Chen, K.-H. and Wallace, D.C. (1992) Southeast Asian Mitochondrial DNA Analysis Reveals Genetic Continuity of Ancient Mongoloid Migrations. *Genetics* **130**: 139-152.
- Bamshad, M., Fraley, A.E., Crawford, M.H., Cann, R.L., Busi, B.R., Naidu, J.M. and Jorde, L.B. (1996) MtDNA Variation in Caste Populations of Andhra Pradesh, India. *Human Biology* **68**: 1-28.
- Bamshad, M.J., Watkins, W.S., Dixon, M.E., Jorde, L.B., Rao, B.B., Naidu, J.M., Prasad, B.V.R., Rasanayagam, A., and Hammer, M.F. (1998a) Female gene flow stratifies Hindu castes. *Nature* **395**: 651-652.
- Bamshad, M., Watkins, W.S., Moore, M.E., Rao, B.B., Naidu, J.M., Prasad, B.V.R., Reddy, P.G., Watkins, C., Rasanayagam, A., Hammer, M.F. and Jorde, L.B. (1998b) MtDNA and Y chromosome variation in South Indian populations. *American Journal of Physical Anthropology* **26(S)**: 27, 67.
- 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.

- Caste (1998) *Microsoft Encarta 98 Encyclopedia*. Redmond, Washington.
- Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: models and estimation procedures. *Evolution* **32**: 550-570.
- Cavalli-Sforza, L.L, Menozzi, P. and Piazza, A. (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton, New Jersey. **212, 239-242, 290-301**.
- Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J. (1988) Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences of the United States of America* **85**: 6002-6006.
- Chen, Y.-S., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A.S. and Wallace, D.C. (1995) Analysis of mtDNA Variation in African Populations Reveals the Most Ancient of All Human Continent-Specific Haplogroups. *American Journal of Human Genetics* **57**: 133-149.
- Cheriyian, C.V. (1973) *A History of Christianity in Kerala*. C.M.S. Press, Kottayam.
- Chu, J.Y., Huang, W., Kuang, S.Q., Wang, J.M., Xu, J.J., Chu, Z.T., Yang, Z.Q., Lin, K.Q., Li, P., Wu, M., Geng, Z.C., Tan, C.C., Du, R.F. and Jin, L. (1998) Genetic relationship of populations in China. *Proceedings of the National Academy of Sciences of the United States of America*. **95**: 11763-11768.
- Cutler, N. (1988) Dravidian Languages and Literatures In: Embree, A.T. (ed) *Encyclopedia of Asian History*. Scriber, New York; Collier Macmillan, London. **1, 399-402**.
- Dobzhansky, T. (1962) *Mankind Evolving*. Yale University Press, New Haven, Connecticut. p. 236.
- Felsenstein, J. (1995). PHYLIP (Phylogeny Inference Package) version 3.57c. Distributed by the author at <http://evolution.genetics.washington.edu/phylip.html> Department of Genetics, University of Washington, Seattle.
- Fitch, W.M. and Margoliash, E. (1967) Construction of Phylogenetic Trees. *Science* **155**: 279-284.
- Ghosh, A. *An Encyclopaedia of Indian Archaeology*. (1989) Munshiram Manoharlal Publishers Pvt. Ltd., New Delhi. **19, 316-317**

- Gibbons, A. Anthropologists Probe Genes, Brains at Annual Meeting. (1998) *Science* **280**: 380-381.
- Grinfeld, A.T. (1998) Tibet In: *Microsoft Encarta 98 Encyclopedia*. Redmond, Washington.
- Gupta, P. S. (ed.) Census of India 1961. (1970) Capital Offset Printers, Delhi. pp. 330-335.
- Hammer, M.F., Karafet, T., Rasanayagam, A., Wood, E.T., Altheide, T.K., Jenkins, T., Griffiths, R.C., Templeton, A.R. and Zegura, S.L. (1998) Out of Africa and Back Again: Nested Cladistic Analysis of Human Y Chromosome Variation. *Molecular Biology and Evolution* **15(4)**: 427-441.
- India. (1998) *Microsoft Encarta 98 Encyclopedia*. Redmond, Washington.
- Kennedy, K.A.R., Deraniyagala, S.U., Roertgen, W.J., Chiment, J. and Sisotell, T. (1987) Upper Pleistocene Fossil Hominids From Sri Lanka. *American Journal of Physical Anthropology* **72**: 441-461.
- Kennedy, K.A.R., Deraniyagala, S.U. (1989) Fossil Remains of 28,000-Year-Old Hominids from Sri Lanka. *Current Anthropology* **30(3)**: 394-399.
- Kennedy, K.A.R., Sonakia, A., Chiment, J. and Verma, K.K. (1991) Is the Narmada Hominid an Indian Homo erectus? *American Journal of Physical Anthropology* **86**:475-496.
- Kogelnik, A.M., Lott, M.T., Brown, M.D., Bavathe, S.B. and Wallace, D.C. (1998) "MITOMAP: a human hitochondrial genome database – 1998 update." *Nucleic Acids Research* **26(1)**: 112-115. <http://www.gen.emory.edu/mitomap.html>
- 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.
- Labie, D., Srinivas, R., Dunda, O., Dode, C., Lapoumeroulie, C., Devi, V., Devi, S., Ramasami, K., Elion, J., Ducrocq, R., Krishnamoorthy, R. and Nagel, R.L. (1989) Haplotypes in Tribal Indians Bearing the Sickle Gene: Evidence for the Unicentric Origin of the β^S Mutation and the Unicentric Origin of the Tribal Populations of India. *Human Biology* **61(4)**: 479-491.

- Lumley, H. de and Sonakia, A. (1985) Contexte stratigraphique et archéologique de l'homme de la Narmada, Hathnora, Madhya Pradesh, Inde. *L'Anthropologie* **89(1)**: 3-12.
- Majumdar, P.P and Mukherjee, B.N. (1993) Genetic diversity and affinities among Indian populations: An overview. In: Majumdar, P.P (ed) *Human Population Genetics*. Plenum, New York. **255-275**.
- Melton, T., Peterson, R., Redd, A.J., Saha, N., Sofro, A.S.M., Martinson, J. and Stoneking, M. (1995) Polynesian Genetic Affinities with Southeast Asian Populations as Identified by mtDNA Analysis. *American Journal of Human Genetics* **57**: 403-414.
- Merriwether, D.A., Hall, W.W., Vahlne, A. and Ferrell, R.E. (1996) mtDNA Variation Indicates Mongolia May Have Been the Source for the Founding Population for the New World. *American Journal of Human Genetics* **59**: 204-212.
- 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.
- Muir, J. (1868) *Original Sanskrit Texts on the Origin and History of the People of India*. Trubner & Co., London, England. pp. 9-11, 140-141.
- Nei, M. (1978) Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics* **89**: 583-590.
- Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York. pp. 176-181.
- Nei, M. and Tajima, F. (1983) Maximum likelihood estimation of the number of nucleotide substitutions from restriction site data. *Genetics* **56**: 979-992.
- Neill, S. (1970) *The Story of the Christian Church in India and Pakistan*. William B Eerdmans Publishing Co., Grand Rapids, Michigan. pp. 16-41.
- Papiha, S.S. (1996) Genetic Variation in India. *Human Biology* **68(5)**:607-628.
- Papiha, S.S., Schanfield, M.S. and Chakraborty, R. (1996) Immunoglobulin allotypes and estimation of genetic admixture among populations of Kinnaur District, Himachal Pradesh, India. *Human Biology* **68**: 777-794.

- 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.
- Reynolds, J. B., B. S. Weir, and C. C. Cockerham. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**: 767-779.
- Roychoudhury, A.K. (1984) Genetic relationships between Indian populations and their neighbours. In:Lukacs J.R. (ed) *The People of South Asia: The Biological Anthropology of India, Pakistan, and Nepal*. Plenum, New York. **283-293**.
- Roychoudhury, A.K. and Nei, M. (1985) Genetic Relationships between Indians and Their Neighboring Populations. *Human Heredity* **35**: 201-206.
- Rychlik, W. and Rhoads, R.E. (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and *in vitro* amplification of DNA. *Nuclie Acids Research* **18**:6409-6412.
- Saitou, N. and Nei, M. (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution* **4**: 406-425.
- Sambrook, J., Fritsch, E.F. and Maniatis, T. *Molecular Cloning: A Laboratory Manual* (2nd ed.). (1989) Cold Spring Harbor Laboratory Press, New York. 6.3-6.19, 14.14-14.19, Appendix B, Appendix E.
- Sankhyan, A.R. (1997) Fossil clavicle of a Middle Pleistocene hominid from the Central Narmada Valley, India. *Journal of Human Evolution* **32**: 3-16.
- Schmidt, K.J. *An atlas and survey of South Asian history*. (1995) M.E. Sharpe, Inc., Armonk, New York.
- Shaffer, J.G. (1984) The Indo-Aryan Invasions: Cultural Myth and Archaeological Reality. In:Lukacs, J.R. (ed) *The People of South Asia: The Biological Anthropology of India, Pakistan, and Nepal*. Plenum, New York. **77-90**.
- Shaffer, J.G. (1998) Indus Valley Civilization. In: *Microsoft Encarta 98 Encyclopedia*. Redmond, WA, USA.
- Sikhs. (1998) *Microsoft Encarta 98 Encyclopedia*. Redmond, Washington.
- Soodyall, H., Vigilant, L., Hill, A.V., Stoneking, M. and Jenkins, T. (1996) mtDNA Control-Region Sequence Variation Suggests Multiple Independent Origins of an

- “Asian-Specific” 9-bp Deletion in Sub-Saharan Africans. *American Journal of Human Genetics* **58**: 595-608.
- Southworth, F.C. (1990) The Reconstruction of Prehistoric South Asian Language Contact. In: Bendix, E.H. (ed) *Annals of the New York Academy of Sciences – The Uses of Linguistics* 583: 207-234.
- Southworth, F.C. (1996) Dravidian Place Names in Maharashtra. In: Houben, J.E.M. (ed) *Indo-Aryan Debate* (collation).
<http://sarsvati.simplenet.com/resources/Indoaryanproblem.htm>
- Spurdle, A., Mitchell, J. and Jenkins, T. (1997) Letters to the Editor. *Human Biology* **69(3)**: 431-435.
- Tambiah, S.J. (1973) From Varna to Caste through Mixed Unions. In: Goody, J. (ed) *The Character of Kinship*. Cambridge University Press, Great Britain. **191-229**.
- Templeton, A.R. (1997) Out of Africa? What do genes tell us? *Current Opinion in Genetics and Development* **7**: 841-847.
- Torrioni, A., Bandelt, H.-J., D’Urbano, L., Lahermo, P., Moral, P., Sellitto, D., Rengo, C., Forster, P., Savontaus, M.-L., Bonn -Tamir, B. and Scozzari, R. (1998) mtDNA Analysis Reveals a Major Late Paleolithic Population Expansion from Southwestern to Northeastern Europe. *American Journal of Human Genetics* **62**: 1137-1152.
- Torrioni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.-L. and Wallace, D.C. (1996) Classification of European mtDNAs From an Analysis of Three European Populations. *Genetics* **144**: 1835-1850.
- Torrioni, A., Lott, M.T., Cabell, M.F., Chen, Y.-S., Lavergne, L. and Wallace, D.C. (1994a) MtDNA and the Origin of Caucasians: Identification of Ancient Caucasian-specific Haplogroups, One of Which Is Prone to a Recurrent Somatic Duplication in the D-Loop Region. *American Journal of Human Genetics* **55**: 760-776.
- Torrioni, A., Miller, J.A., Moore, L.G., Zamudio, S., Zhuang, J., Droma, T. and Wallace, D.C. (1994b) Mitochondrial DNA Analysis in Tibet: Implications for the Origin of the Tibetan Population and Its Adaptation to High Altitude. *American Journal of Physical Anthropology* **93**: 189-199.

- Torrioni, A., Neel, J.V., Barrantes, R., Schurr, T.G. and Wallace, D.C. (1994c) Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. *Proceedings of the National Academy of Sciences of the United States of America* **91**: 1158-1162.
- Torrioni, A., Petrozzi, M., D'Urbano, L., Sellitto, D., Zeviani, M., Carrara, F., Carducci, C., Leuzzi, V., Carelli, V., Barboni, P., De Negri, A. and Scozzari, R. (1997) Haplotype and Phylogenetic Analyses Suggest that One European-Specific mtDNA Background Plays a Role in the Expression of Leber Hereditary Optic Neuropathy by Increasing the Penetrance of the Primary Mutations 11778 and 14484. *American Journal of Human Genetics* **60**: 1107-1121.
- Torrioni, A., Schurr, T.G., Cabell, M.F., Brown, M.D., Neel, J.V., Larsen, M., Smith, D.G., Vullo, C.M. and Wallace, D.C. (1993a) Asian Affinities and Continental Radiation of the Four Founding Native American mtDNAs. *American Journal of Human Genetics* **53**: 563-590.
- Torrioni, A., Sukernik, R.I., Schurr, T.G., Starikovskaya, Y.B., Cabell, M.F., Crawford, M.H., Comuzzie, A.G. and Wallace, D.C. (1993b) mtDNA Variation of Aboriginal Siberians Reveals Distinct Genetic Affinities with Native Americans. *American Journal of Human Genetics* **53**: 591-608.
- Wallia, C.J.S. (1996) *Ancient India in a New Light*.
<http://www.indiastar.com/ancient.htm>
- Zaykin, D. and Pudovkin, A. (1993) 2 Programs to Estimate Significance of Chi-2 Values Using Pseudo-Probability Tests. *Journal of Heredity* **84(2)**:152.