mtDNA Variation and Human Population Structure in India
MITOCHONDRIAL DNA SEQUENCE VARIATION
AND
HUMAN POPULATION STRUCTURE
IN THE
INDIAN SUBCONTINENT

By
Anouk Margareta Pramila Behara, 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
© Anouk Behara, October 1995
TITLE: Mitochondrial DNA Sequence Variation and Human Population Structure in the Indian Subcontinent

AUTHOR: Anouk Margareta Pramila Behara, B.Sc. (McMaster University)

SUPERVISOR: Professor R.S. Singh

NUMBER OF PAGES: xi, 163
Abstract

An analysis of mitochondrial DNA (mtDNA) sequence variation among individuals of Indian origin was performed to compare mtDNA diversity within the Indian population to that of other human populations, and to examine Indian population substructure related to geographical location, spoken language and the caste system. A 195 bp non-coding segment of mitochondrial DNA was amplified and sequenced from 84 individuals representing a broad geographical distribution and three of the major castes. Sixty-one different haplotypes, and 54 polymorphic sites, corresponding to 28% of the sites analyzed, were found. The results obtained from the analysis of distributions of pairwise sequence differences, as well as phylogenetic analysis, are in agreement with rapid population expansion or steady growth as opposed to the maintenance of a constant population size. At 4 out of the 54 variable sites detected there was a significant difference between northern and southern populations. However, phylogenetic analysis revealed no significant population substructure with respect to either geographical distribution, language or the caste system. Some minor effects were observed in a significant under-representation of Dravidian language speakers, corresponding to individuals of south Indian origin, within one cluster, and a similar under-representation of members of the Vaisya caste in another cluster. The
phylogenetic tree obtained from integration of the Indian sequences with sequences from other human populations produced a starlike cluster encompassing both Indian and non-Indian individuals. Clustering of Indians within the tree was not completely random as there were significantly fewer Indians than expected within one cluster. The results suggest that the time elapsed since the origin of the caste system has not been sufficiently long to produce significant differentiation between the major castes. Alternatively, the mtDNA fragment chosen for analysis may contain too few major polymorphisms to balance out background noise, and is therefore not adequately sensitive to detect local population structure.
With my best suit on and a smile,
I dedicate this thesis in fond memory to my father,

Dr. Minaketa Behara

without whom I would never have learned to believe that your only limitations
are those that you set for yourself, and who, by example, taught me how to
keep on going in the face of all adversity.
Acknowledgements

A great deal of collaborative effort goes into a project such as a master's thesis, and thanks are due to the many people who helped along the way in a myriad of fashions. Firstly I would like to thank my teachers: Dr. Rama Singh for his support and guidance over the last few years, and Dr. Richard Morton for many good suggestions and valuable critiques. I would also like to thank Dr. Brian Golding for his advice on statistical and UNIX problems.

For the often tedious task of obtaining enough blood samples for this study, the following individuals deserve recognition: Suman Mukhopadhyay, and Senthil Muthuswamy for their willingness to help in making contacts to obtain samples; Lindsay Smith for taking blood samples for me, often on very short notice; and Dr. Patricia Chang and Rebecca Ott for supplying my first set of samples, and for helping me set up contacts.

To the members of Life Sciences 505, past and present, abundant thanks are owed. Specifically, to Fayaz Alladina, for showing me around, and for being willing to share with me his wealth of knowledge on molecular biology techniques, all while he had a thesis of his own to finish. Francisco Molina for being all too willing to drop everything to help me out with my many, many questions. Ling-Wen Zeng for additional help and philosophical discussions (and table tennis). Tony Long, without whom I would not be here in the first place, for always forcing me to think about what I was doing, and why I was doing it. Special thanks are owed to my lab-mates and friends Alberto Civetta and Rob Kulathinal, for ideas,
discussions, a helping hand . . . and for Spanish and soccer lessons.

Thanks also to Daniel Fieldhouse for always being willing to take time out help out and advise (and lend equipment to) that new kid who didn't know a thing about the art of sequencing. A good deal of gratitude is owed to Fariborz Yazdani, for his generous help with PCR, sequencing, and of course anything to do with UNIX, and for laughing with me whenever I discovered yet another novel way to mess things up. There are countless others who helped in out in many a panic, freely lending out knowledge and equipment, without whom my work would have been impossible.

I would like to express my appreciation to the "5th floor gang" (including a few honorary members) for good company, good food, and most of all, for many fond memories (and loaned Nighthawks). Special thanks go to my husband Ken for always helping out, sharing his vast computing knowledge, making many good suggestions, but most importantly for just putting up with me and being there for me during this phase of my life. Last, but by no means least, a most sincere thank you goes out to my family for their support, love, and most of all, their belief in me.
# Table of Contents

**Introduction** .................................................. 1  
The Indian Subcontinent ........................................ 1  
Historical Overview ............................................. 1  
The Languages of India ........................................... 9  
Religions and the Caste System ................................ 13  
Genetic Surveys of the Indian Population ...................... 18  
MtDNA Sequence Analysis in Human Population Studies ....... 24  
Previous Human Population Studies Based on mtDNA ........ 26  
Objectives of the Present Study ................................ 29

**Materials and Methods** ........................................ 32  
Sampling procedure ............................................... 32  
DNA Extraction ..................................................... 32  
PCR Amplification of 1024 bp Region of Mitochondrial Genome 37  
Sequencing of 195 bp Region of Mitochondrial D-loop ........ 40  
Sequence Analyses ................................................. 42  
Phylogenetic Analysis ............................................ 42  
Distributions of Frequencies of Pairwise Differences ........ 47  
Polymorphic Sites ................................................ 50

**Results** .......................................................... 52  
General Summary of Sequence Analysis ......................... 52  
Phylogenetic Analysis ............................................ 60  
Distributions of Pairwise Sequence Differences .............. 81  
Polymorphic Sites ................................................. 86

**Discussion** ...................................................... 115  
Phylogenetic Inferences ......................................... 115  
Population Demography ........................................... 121  
Informative Polymorphic Sites ................................ 123  
Indians Within the Global Human Population .................. 125  
New directions for Future Research ............................ 125

**Conclusion** ..................................................... 128
List of Tables

Table 1: Proportions of religions in India ............................................. 16
Table 2: Summary of previous genetic studies on the Indian population ........................................ 24
Table 3: Geographical distribution of individuals ........................................ 35
Table 4: Distribution of languages spoken by individuals ........................................ 36
Table 5: Distribution of religions and castes for individuals ........................................ 36
Table 6: Origins of sequences used in phylogenies ........................................ 46
Table 7: Sequence variability within the Indian population ........................................ 53
Table 8: Definitions of sample code abbreviations ........................................ 58
Table 9: Heterozygosities for different castes and religions ........................................ 59
Table 10: Heterozygosities for northern vs. southern Indians ........................................ 59
Table 11: Steps required for different tree-inference methods ........................................ 61
Table 12: Mean sequence differences within different castes and religions ........................................ 82
Table 13: Test of Goodness-of-fits to Poisson distribution ........................................ 87
Table 14: Chi-square analysis of substitution frequencies in southern vs. northern Indian individuals ........................................ 88
Table 15: Frequencies of T→C substitution at polymorphic site 16126 ........................................ 93
Table 16: Frequencies of G→A substitution at polymorphic site 16129 ........................................ 96
Table 17: Frequencies of T→C substitution at polymorphic site 16172 ........................................ 99
Table 18: Frequencies of C→T substitution at polymorphic site 16223 ........................................ 102
Table 19: Frequencies of T→C substitution at polymorphic site 16304 ........................................ 105
Table 20: Frequencies of G→A substitution at polymorphic site 16309 ........................................ 108
Table 21: Frequencies of T→C substitution at polymorphic site 16311 ........................................ 111
Table 22: Frequencies of A→T substitution at polymorphic site 16318 ........................................ 114
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1:</td>
<td>Map of India and surrounding countries</td>
<td>3</td>
</tr>
<tr>
<td>Figure 2:</td>
<td>Distribution of spoken languages in India</td>
<td>15</td>
</tr>
<tr>
<td>Figure 3:</td>
<td>Geographic distribution of samples</td>
<td>34</td>
</tr>
<tr>
<td>Figure 4:</td>
<td>Reference sequence of mitochondrial D-loop</td>
<td>39</td>
</tr>
<tr>
<td>Figure 5:</td>
<td>Image of autoradiograph of sequence</td>
<td>44</td>
</tr>
<tr>
<td>Figure 6:</td>
<td>Geographical regions of India</td>
<td>49</td>
</tr>
<tr>
<td>Figure 7:</td>
<td>Variable sites for 84 mitochondrial D-loop sequences</td>
<td>55</td>
</tr>
<tr>
<td>Figure 8:</td>
<td>UPGMA tree of all 84 sequences</td>
<td>57</td>
</tr>
<tr>
<td>Figure 9:</td>
<td>Phylogenetic tree - 84 Indian individuals</td>
<td>64</td>
</tr>
<tr>
<td>Figure 10:</td>
<td>Phylogenetic tree - 84 Indian individuals, based on geography</td>
<td>66</td>
</tr>
<tr>
<td>Figure 11:</td>
<td>Phylogenetic tree - 162 Indian individuals, based on geography</td>
<td>68</td>
</tr>
<tr>
<td>Figure 12:</td>
<td>Phylogenetic tree - 78 Indian individuals, based on language</td>
<td>70</td>
</tr>
<tr>
<td>Figure 13:</td>
<td>Phylogenetic tree - 71 Indian individuals, based on caste</td>
<td>73</td>
</tr>
<tr>
<td>Figure 14:</td>
<td>Phylogenetic tree - 149 Indian individuals, based on caste</td>
<td>75</td>
</tr>
<tr>
<td>Figure 15:</td>
<td>Phylogenetic tree - North Indians only, based on caste</td>
<td>78</td>
</tr>
<tr>
<td>Figure 16:</td>
<td>Phylogenetic tree - Indians and other human populations</td>
<td>80</td>
</tr>
<tr>
<td>Figure 17:</td>
<td>Distributions of frequencies of pairwise sequence differences</td>
<td>85</td>
</tr>
<tr>
<td>Figure 18:</td>
<td>Geographic distribution of substitutions at site 16126</td>
<td>92</td>
</tr>
<tr>
<td>Figure 19:</td>
<td>Geographic distribution of substitutions at site 16129</td>
<td>95</td>
</tr>
<tr>
<td>Figure 20:</td>
<td>Geographic distribution of substitutions at site 16172</td>
<td>98</td>
</tr>
<tr>
<td>Figure 21:</td>
<td>Geographic distribution of substitutions at site 16223</td>
<td>101</td>
</tr>
<tr>
<td>Figure 22:</td>
<td>Geographic distribution of substitutions at site 16304</td>
<td>104</td>
</tr>
<tr>
<td>Figure 23:</td>
<td>Geographic distribution of substitutions at site 16309</td>
<td>107</td>
</tr>
<tr>
<td>Figure 24:</td>
<td>Geographic distribution of substitutions at site 16311</td>
<td>110</td>
</tr>
<tr>
<td>Figure 25:</td>
<td>Geographic distribution of substitutions at site 16318</td>
<td>113</td>
</tr>
</tbody>
</table>
Introduction

The Indian Subcontinent

The human population of the Indian subcontinent displays some of the greatest diversity with respect to the physical appearance, spoken languages, religious and cultural traditions and genetic diversity of its inhabitants. The country of India has a population of over 800 million inhabitants (Cavalli-Sforza et al., 1994). See figure 1 for a current map of India and surrounding the countries.

The genetic diversity seen within India may be due to a number of different factors, including the large number of migrations and invasions into India over the last 4000 or so years, most of which have left a cultural, linguistic, and presumably genetic, signature on its inhabitants. Some of the diversity may also be the result of the stratified social system practiced by Hindus in India, who make up approximately 83% of the population. This system, known as the caste system, applies strict rules to its members regarding restrictions on marriage between members of different castes.

Historical Overview

Information regarding the very original inhabitants of India is sparse. Numerous theories exist addressing the origins of the Dravidian peoples, as well
Figure 1: Map of India and surrounding countries.
India and Surrounding Countries

1 Sikkim
2 Arunachal Pradesh
3 Nagaland
4 Manipur
5 Mizoram
6 Tripura
7 Meghalaya
as of the tribal populations, which are generally believed to be descendants of the original inhabitants of the subcontinent. As historical accounts are sparse prior to a few centuries before Christ, most theories are based on archeological and linguistic evidence.

**Tribal Populations**

Present day tribal populations are considered by many scholars to be descendants of the very original inhabitants of the Indian subcontinent. Their origins have been postulated as being Negroid, followed by Proto-Australoid, then Mongoloid. Evidence for a Palaeolithic Australoid occupation includes an Australoid component which can be seen in some present day tribals, along with the Austro-Asiatic languages spoken by many tribals. A Dravidian origin for some tribal populations has also been suggested. The Aryan invaders are believed to have found several different tribes, belonging to two main types of people, on arriving in India. These were the Dravidians and the Kolorians, an Austro-Asiatic people, supposed to be the ancestors of the present day Munda and Santhal tribes. The Kolorians are thought to have arrived in India prior to the Dravidians (Cavalli-Sforza *et al.*, 1994; Guha, 1944; Vidyarthi and Rai, 1976).

**Harappan Culture**

The first major civilization in India is believed to have been the Harappan civilization of the Indus valley. This urban civilization was based in the cities of Mohenjo-Daro and Harappa, located near what is now the city of Lahore in Pakistan. Radiocarbon dating of archeological remains at these and more than
200 other Harappan sites indicate that the culture lasted from about 2300-1750 B.C. (Agrawal, 1971). The culture was highly developed, with organized cities, and was based on agriculture.

Biometric measurements and comparisons of cranial series of Harappans to series from sites dated to about the same time reveal similarities between the Harappans and Tepe Hissar and Sakkara civilizations, which were based in Iran and in Egypt respectively, suggesting a possible relation to Mesopotamian civilizations.

There is archeological evidence of a pre-Harappan culture, which, rather than being replaced by, shows signs of having possibly evolved into the Harappan culture.

The seemingly sudden disappearance of the Harappan civilization between 1750 and 1500 B.C. remains a mystery. Speculations based on the comparison of Harappan skeletal remains, and modern crania from the same region suggest that there may be a degree of genetic continuity between Harappans and the present inhabitants of the region. Other theories suggest that the Harappans moved southwards, or that their reign ended due to natural disaster. The most prevalent theory is that the civilization saw its end as a consequence of the Aryan conquest (Cavalli-Sforza et al., 1994; Dutta, 1984; Sherratt, 1980).

The question remains as to whether the Harappans were the ancestors of the present day Dravidians. Scripts have been found at Harappan sites, however they cannot be deciphered, and it is unclear as to whether they are written in a
Dravidian language or not.

Dravidians

The origin of the Dravidian people is under considerable debate. Accounts from the Vedic literature, a collection of texts describing Hindu society and history which were originally handed down orally, speak of the Dasyu, a population of dark people encountered by the Aryans on their arrival in India. It is believed that these people may have been Dravidians.

It has been speculated that the Dravidian civilization may well have arisen out of the Harappan culture. Linguistic evidence does suggest that the Dravidian culture was previously spread out over the whole of India. As well, speakers of the Brahui Dravidian language in Baluchistan consider the inhabitants of Mohenjo-Daro to be their ancestors (Sinha, 1977). This still does not answer the question of the origin of the Harappan culture. Whether the original Dravidian culture spread from north to south or from south to north remains under debate, although geological evidence supports the latter, as the Himalayan foothill region of India was under water up until 5000-7000 years ago (Sinha, 1977).

It has also been suggested that the Dravidians spread from Iran down into India (Sinha, 1977). Cavalli-Sforza (1994) suggests that this occurred in the form of a migration of early Neolithic farmers. Archeological evidence exists of agriculture in India as early as 5500 B.C. A Mediterranean origin of Dravidians has also been postulated (Guha, 1944).
Aryan Invasions

Around 3500 years ago saw the beginnings of the Aryan invasions into India. Evidence suggests that these invaders may have belonged to a group of people known as the Kurgans. The Kurgans are postulated to have been a semi-nomadic population, either pastoral nomads or mounted warriors, originating around 4000 B.C. in the Pontic Steppes, located in what is now the Ukraine. The Kurgans used iron and are believed to have been the first to harness and ride the horse 5000 years ago, and with their newfound ease of transport, spread first to the Danube river area around 3500 B.C., and then eastwards, eventually arriving in the Punjab around 1500 B.C.. From there, they continued to spread out gradually over northern India, reaching Bengal in around 800 B.C. (Anthony et al., 1991; Cavalli-Sforza et al., 1994; Cayne, 1988).

Whatever their origin, the Aryan invaders are believed to have introduced the Vedic religion, which was later to become the Hindu religion, as well as the concept of varnas, which later crystallized into the caste system.

Later Invasions

Following the Aryan invasion, a number of smaller invasions took place in India, all or some of which may have contributed in some way to the diversity found within India.

Alexander the Great's expeditions included a sojourn in India in around 325 B.C., with the only major impact appearing to have been the establishment of Greek style cities. Shortly after Alexander's campaign, the Mauryan empire in
India was established. This empire lasted about 100 years, and included almost the entire subcontinent of India, excepting the very southern part. The language spoken by this civilization belonged to the Indo-Aryan family. It was during the reign of Asoka, the third ruler of the Maurya line, from 273-232 B.C., that Buddhism rose in popularity, replacing the former state religion, Hinduism. Following the decline of the Mauryan empire, numerous small kingdoms existed in India. Northern India, especially Punjab, found itself under numerous different rulers, originating mainly from central Asia, including the Bactrian Greeks, who founded the Indo-Greek dynasty in the Ganges valley, around the 2nd century B.C. Rule of northern India passed from the Parthians to the Scythians, who managed to dislodge the last of the Indo-Greek empire. The resulting Indo-Scythian culture created the Kushan dynasty, which dominated central Asia and northern India. Around 318 A.D. saw the birth of the Gupta Dynasty, which lasted until 544 A.D., and which fended off attempted invasions from Iranians, as well as of those Mongoloid people known as the White Huns, who are believed to have been related to Turkic people.

Later Hunnish invasions were more successful, resulting in a short reign in Kashmir. An Arabic invasion occurred in the 7th and 8th centuries, but did not lead to a prolonged rule. Northern India was under Turkish influence starting in the 11th century. The introduction of Islam into India as a major religion occurred around this time. The Mogul empire was founded by Babur in 1526 and lasted 250 years, encompassing all of northern India. The decline of the Mogul empire
was witness to yet another Persian invasion.

It appears that the two major foreign influences on ancient India, at least in the north, were of Indo-European or Mongoloid origin, usually in the form of invasions entering India from the north-west. There has also been a more or less continuous infiltration of people of Mongolian descent through the Himalayan and Burmese valleys.

In more recent history, a number of European settlements occurred in India. These began with the Portuguese in the early 16th century, who settled mainly in Goa. Later settlers included the Dutch, the French and of course the British, who established colonial rule of India from late 18th century until 1947 (Cavalli-Sforza et al., 1994; Cayne, 1988; Masson-Oursel et al., 1967; Wells, 1971).

The Languages of India

Genetic differences between closely related individuals have been shown to exist in Europe on the basis of spoken language. Barbujani (1990) reported that zones of sharp changes in gene frequencies corresponded with linguistic boundaries. In general, it appears that distributions of genes correspond very well with language distributions for the majority of human populations (Cavalli-Sforza, 1991).

Native language is therefore addressed in this study as a potential factor influencing genetic diversity. A large number of different languages are spoken in India. The majority of these fall into two major language families, the
Indo-European family, and the Dravidian family, where a language family is defined as a group of languages for whom a historical relationship of some sort is probable. Languages belonging to a third language family, the Austro-Asiatic family, are also spoken by a significant number of individuals.

**Austro-Asiatic Language Family**

The language spoken by the original inhabitants of India is unknown. A precursor to the Munda group of languages is a possibility. Alternatively, they may have been a much more recent introduction into the linguistic tapestry of India, as a consequence of the many migrations or invasions originating from the east or northeast of India.

The Munda languages are spoken by over 6 million Indian individuals, mainly in the north-eastern sections of India. The major representatives of this group are Mundari, spoken in Madhya Pradesh, Orissa and Bihar, and Santali, spoken in Orissa, Bihar and West Bengal (Crystal, 1987).

**Dravidian Language Family**

The Dravidian language family comprises over 20 different languages, which are most highly localized in the southern part of India, namely Karnataka, Kerala, Tamil Nadu and Andhra Pradesh, as well as in the northern part of Sri Lanka. Small pockets of Dravidian language speakers also exist in areas of India further north. The major examples are Gondi in the southern parts of Madhya Pradesh, Konda and Kui in Orissa, Kurukhi in north-eastern India, Malto in north-eastern Bihar and, 1000 miles north of the majority of Dravidian language speakers, Brahui
in Baluchistan (Pakistan). Dravidian language speakers number approximately 150 million.

Little information is available concerning the origins of either the language or its speakers. Suggestions include that the original speakers migrated from lands to the South, which are now submerged, or that they moved in from the north-west, originating in Asia around 4000 B.C.. Another theory suggests that the proto-Dravidian languages arrived first in Iran, and then moved to India along with early Neolithic farmers (Cavalli-Sforza et al., 1994; Renfrew, 1989). A controversial relationship with the Uralic and Altaic languages, spoken in northern Europe and Russia, and central Russia and Mongolia respectively, has also been proposed (Cavalli-Sforza et al., 1994; Crystal, 1987).

There is support for the idea that Dravidian languages were once spoken in northern India as well, and were then gradually displaced or suppressed by the languages spoken by the Indo-European invaders. The wide distribution of pockets of Dravidian languages found over the northern region of India suggests that these may be remnants of a much wider distribution of Dravidian languages over the entire subcontinent (Crystal, 1987).

**Indo-European Language Family**

Languages belonging to the Indo-European language family are spoken by almost half of the world's population. The majority of the languages spoken by non-tribal individuals in northern India belong to the branch of Indo-European languages known as the Indo-Iranian languages. Within this branch, they fall into
the subgroup known as Indo-Aryan.

As with the origin of the Dravidian languages, the origins of the Indo-European languages in India are not entirely certain. The Indo-European language family is generally thought to have its origins in the steppes of Russia, where Proto-Indo-European is believed to have been spoken by the Kurgans. It would then have spread first to Europe and later to India along with the Kurgan migrations. The introduction of the language into India by the Aryans would have begun in around 1400 B.C..

Indo-European languages are believed to have gradually replaced the previously much more widely distributed Dravidian languages in northern India (Cavalli-Sforza et al., 1994; Trautmann, 1981).

Language Interrelationships

When postulating the replacement of one language by another, it must be recalled that languages continually evolve, and that the replacement of one language by another may not be complete due to the influence of the original regional language on it.

Indo-European based languages spoken in northern India show some evidence of common development along with the Dravidian languages (Renfrew, 1987). Evidence of this is seen in what are called loan words, along with some shared grammatical structure. This could imply a shared period of cohabitation of the two cultures, presumably leading to some degree of some gene flow along with semantic flow between the populations.
Parallels between linguistics and origins of populations should be drawn with caution, recalling that language patterns can evolve without any corresponding movement of peoples. The wave hypothesis of language spread, proposed by Johannes Schmidt in 1872, suggests that linguistic changes spread over speech areas in waves that do not necessarily involve any movement of peoples, but may rather be the result of dominant political or economic influences (Renfrew, 1987).

The map of India and surrounding countries shown in figure 2 describes what languages are spoken by the majority of the inhabitants of each province.

**Religions and the Caste System**

As with languages and appearances, the religious beliefs of Indians are quite diverse. Relative distributions of religions have remained more or less unchanged over the last 5 centuries (Cassero and Modiano, 1993). Conversions to Christianity and Islam were most prevalent during the 15th and 16th centuries, and decreased abruptly thereafter. Conversion from Christianity or Islam to Hinduism is quite uncommon (Cassero and Modiano, 1993). Table 1 lists the proportions of members of the major religious groups of India. According to the census, Hinduism is by far the most common religion in India (83%). It should be kept in mind however, that tribals (who are mainly animists) and atheists have traditionally been classified as Hindus in censuses, and therefore the proportion of Hindus is in all likelihood over-represented.
Figure 2: Distribution of spoken languages in India.

Map of India showing distribution of the languages spoken in India. Shaded regions correspond to the language family to which the language spoken by the majority of the region's inhabitants belongs. Map reproduced from Singh, 1993.
INDIA
Distribution of Language Families and Major Languages

Andamanese
Austro-Asiatic
Dravidian
Indo-Aryan
Dardic Group
Tibeto-Burman

Scale: 1:14,000,000 Or 1 Cm. = 140 Kms.
Table 1: Proportions of religions in India. These percentages are taken from the 1981 Census of India.\(^a\)

<table>
<thead>
<tr>
<th>Religion</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindu</td>
<td>82.60 %</td>
</tr>
<tr>
<td>Muslim</td>
<td>11.40 %</td>
</tr>
<tr>
<td>Christian</td>
<td>2.40 %</td>
</tr>
<tr>
<td>Sikh</td>
<td>2.00 %</td>
</tr>
<tr>
<td>Buddhist</td>
<td>0.71 %</td>
</tr>
<tr>
<td>Jain</td>
<td>0.48 %</td>
</tr>
<tr>
<td>Other</td>
<td>0.42 %</td>
</tr>
<tr>
<td>Not Stated</td>
<td>0.01 %</td>
</tr>
</tbody>
</table>

\(^a\) Cassero and Modiano, 1993.
The Caste System

Within the Hindu religion exists a system of social stratification known as the caste system. Castes are endogamous groups, comprising individuals of the same occupations. The term caste includes both the concepts of varna (meaning colour) and jati (referring to birth). Jati is the more precise term for occupational group, as a person belongs to his jati from birth onwards, and may not change occupations. Early accounts found in the Veda suggest that the caste system was introduced to India by the Aryans at their conquest.

The word 'caste' traces its origin to the Portuguese 'casta', meaning 'race' (Masson-Oursel et al., 1967), suggesting that the subdivision of individuals into the various castes was based on physical appearance and/or ethnic heritage. This idea is supported by the fact that the word varna means colour, again suggesting that the original subdivision of individuals into different castes may have been based on skin colour. If this is in fact the case, genetic differences should exist between members of different castes, as a result of disparate ethnic heritages. Some theories propose that the Aryans subdivided Dravidians into lower castes. Given the theoretically strict endogamy within castes, there should be little genetic flow between members of different castes. Mixture between castes however, is not unheard of or even uncommon, with offspring from such marriages being placed in the lower caste (Masson-Oursel et al., 1967).

Although there were originally four castes, the number of castes has grown tremendously since the inception of the system. Every region has many different
castes, many of which are region-specific (Masson-Oursel et al., 1967). At present there are five main castes, with the inclusion of the so-called Scheduled Castes. In order of status, the castes are: The Brahmins (priests), the Kshatriyas (warriors), the Vaisyas (businessmen), the Sudras (labourers), and the Scheduled Castes. The Sudra caste may also not have been in existence from the start, as there are no accounts of this caste in the Vedic literature (Masson-Oursel et al., 1967).

Genetic Surveys of the Indian Population

Previous studies addressing genetic variation in India have mainly been based on the use of gene frequency data, employing markers such as immunoglobulins, MHC types, and blood group frequencies. Results from these studies vary with respect to geographic and/or caste variation.

A north-south genetic difference has been observed by Schanfield (1981) on the basis of immunoglobulin allotypes. The genetic difference observed was larger between high castes in the north and low castes in the south.

Castes

A number of studies addressing caste differentiation have been performed on various regions of India. Studies based on the use of different genetic markers yield disparate results, where some markers exhibit significant differences between castes while others do not. Table 2 summarizes the genetic markers used in these studies.
<table>
<thead>
<tr>
<th>Study</th>
<th>Factors studied</th>
<th>Region</th>
<th>Markers</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schanfield and Kirk (1981)</td>
<td>geography/caste</td>
<td>Various</td>
<td>Km1, A2m</td>
<td>no, yes</td>
</tr>
<tr>
<td>Kamboh (1984)</td>
<td>caste</td>
<td>North India</td>
<td>Pl, TF, GC, PGM1</td>
<td>yes</td>
</tr>
<tr>
<td>Das et al. (1986)</td>
<td>caste</td>
<td>Assam</td>
<td>ABO blood groups</td>
<td>yes</td>
</tr>
<tr>
<td>Malhotra et al. (1978)</td>
<td>caste</td>
<td>Maharashtra</td>
<td>Serological/biochemical loci</td>
<td>yes</td>
</tr>
<tr>
<td>Walter et al. (1993)</td>
<td>caste</td>
<td>Andhra Pradesh</td>
<td>HP, TF, GC, PI</td>
<td>no, yes, yes, yes</td>
</tr>
<tr>
<td>Papiha et al. (1989)</td>
<td>religion</td>
<td>U.P./Andhra Pradesh</td>
<td>HLA Antigens</td>
<td>yes</td>
</tr>
<tr>
<td>Roychoudhury (1974)</td>
<td>caste</td>
<td>North India</td>
<td>Protein/enzyme loci</td>
<td>no</td>
</tr>
<tr>
<td>Sunderland et al. (1976)</td>
<td>caste</td>
<td>North India</td>
<td>Haptoglobins</td>
<td>no</td>
</tr>
<tr>
<td>Chahal et al. (1986)</td>
<td>caste</td>
<td>Punjab</td>
<td>GLO1</td>
<td>no</td>
</tr>
<tr>
<td>Kushwaha et al. (1990)</td>
<td>caste</td>
<td>Haryana</td>
<td>Blood groups, protein/enzyme loci</td>
<td>no</td>
</tr>
<tr>
<td>Chakraborty et al. (1977)</td>
<td>caste</td>
<td>Maharashtra</td>
<td>Serological/biochemical loci</td>
<td>no</td>
</tr>
<tr>
<td>Ghosh et al. (1981)</td>
<td>caste</td>
<td>West Bengal</td>
<td>G6PD, haemoglobin loci</td>
<td>no</td>
</tr>
<tr>
<td>Chakraborty et al. (1987)</td>
<td>caste</td>
<td>West Bengal</td>
<td>Gm, Km, IgG</td>
<td>no</td>
</tr>
<tr>
<td>Chakraborty et al. (1975)</td>
<td>caste</td>
<td>West Bengal</td>
<td>Blood groups</td>
<td>no</td>
</tr>
<tr>
<td>Chakraborty et al. (1986)</td>
<td>caste</td>
<td>West Bengal</td>
<td>Various markers</td>
<td>no</td>
</tr>
<tr>
<td>Sethuraman et al. (1982)</td>
<td>caste</td>
<td>Andhra Pradesh</td>
<td>MN Blood groups</td>
<td>no</td>
</tr>
<tr>
<td>Rajanimumari et al. (1991)</td>
<td>caste</td>
<td>Andhra Pradesh</td>
<td>ABO system</td>
<td>no</td>
</tr>
<tr>
<td>Rajeskar et al. (1987)</td>
<td>caste</td>
<td>Tamil Nadu</td>
<td>HLA Antigens</td>
<td>no</td>
</tr>
<tr>
<td>Saha et al. (1992)</td>
<td>caste</td>
<td>Bihar</td>
<td>TF, GC, PGM, GLO1, PGD, AK</td>
<td>no</td>
</tr>
</tbody>
</table>

Abbreviations:
- AK: Adenylate Kinase
- Am: Alpha-heavy chain of immunoglobulin IgA
- G6PD: Glucosephosphate Dehydrogenase
- GC: Group-specific component
- GLO: Glycoxylase-I
- Gm: Gamma-light chain of immunoglobulin IgG
- HP: Haptoglobin
- IgG: Immunoglobulin
- Km: Kappa-light chain of immunoglobulin IgG
- PGD: Phosphogluconate Dehydrogenase
- PGM: Red-cell phosphoglucomutase
- PI: α1-antitrypsin
- TF: Transferrin
In one study on north Indians, Brahmins and Kshatriyas clustered together, away from Vaisyas and members of the Scheduled Castes (Kamboh, 1984). Frequency differences between four castes in Assam were mostly statistically significant (Das et al., 1986). Serological allele frequencies showed a small degree of genetic differentiation between castes in Maharashtra, suggesting that gene-flow among castes has been small in the recent past. The amount of differentiation was small, however, implying that the amount of time to achieve a significant degree of intercaste variation has not been sufficient since the split into castes (Malhotra et al., 1978). Intercaste variation was statistically significant for 3 out of 4 different markers in Andhra Pradesh (Walter et al., 1993). A statistically significant amount of variation was also observed for certain HLA Antigens between Hindus from Uttar Pradesh and Muslims from Andhra Pradesh and Gujarat. Whether these differences are the result of religious differentiation or of geographical effects can not be determined (Papiha et al., 1989).

Numerous studies have found no differentiation between castes. A study on north Indians using protein loci, and addressing both caste and linguistic groups, showed that only 1% of total genetic diversity can be attributed to between population diversity as opposed to within population diversity. This has been ascribed to gene flow between populations (Roychoudhury, 1974). No statistically significant differences were found between castes in various studies on north Indians (Chahal et al., 1986; Kushwaha et al., 1990; Sunderland et al., 1976). A low, but not significant, amount of intercaste differentiation was observed in
Maharashtra (Chakraborty et al., 1977). Three studies in West Bengal revealed no intercaste differences (Chakraborty et al., 1975; Chakraborty et al., 1987; Ghosh et al., 1981). No differences were seen in a study using both high and low castes, implying that the high castes were closely related to low ranked scheduled castes in this area (Chakraborty et al., 1986). Three studies on south Indians, from Andhra Pradesh and Tamil Nadu, also showed no significant differences, and again demonstrated that most of the genetic diversity found was within populations, rather than between populations (Rajanikumari and Srisailapathy, 1991; Rajasekar et al., 1987; Sethuraman et al., 1982).

Cavalli-Sforza (1994) used genetic frequency markers to address the question of isolation of only the Brahmin caste from other castes. Six geographically disparate populations of Brahmins did not cluster with each other. This suggests that any historical genetic relationship between them may have been obliterated by geographic isolation and subsequent genetic drift.

A study conducted on Bihar populations, using allele frequencies, demonstrated that Brahmins and Muslims clustered together, separately from low caste and tribal samples (Saha et al., 1992). The average heterozygosities computed for the Brahmin and the Muslim groups were also lower.

Tribals

Tribal populations have been shown to be genetically distinct from non-tribal populations in south India (Kamboh, 1984). The non-tribal south Indians cluster more closely with north Indian non-tribal caste populations, implying a greater
degree of genetic relation with geographically distant members of the caste society than with proximal members of tribal communities, thus reinforcing the idea of a common origin for members of the caste community, distinct from that of tribal Indians.

Two groups of tribals also clustered in a very distinct cluster in a study in Assam. One group of tribals however, clustered with groups believed to be of Mongolian descent, and was closer to groups of Hindu caste members than to the other two tribals groups (Walter et al., 1986). These tribals may not trace their ancestry to the original inhabitants of India, or there may have been significant amounts of gene flow between these tribals and more recent inhabitants of the area.

A north-south cline has been observed by Chahal (1986), for tribal groups. The north-south gradient is not seen among caste groups, in fact the opposite trend is seen (Mukherjee et al., 1986). A unicentric origin has also been proposed for tribal groups based on distributions of the sickle gene mutation, which is postulated to have arisen prior to the dispersion of tribal groups (Labie et al., 1989).

Language

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

Clustering on the basis of language was also demonstrated in a study in Bihar, where Brahmins and Muslims (both speaking Indo-European languages), formed one cluster, while a second cluster contained one subcluster consisting of Mundari speakers (an Austro-Asiatic language) and one subcluster consisting of speakers of Dravidian language family.

A study conducted in Assam however, had one cluster containing Indo-European languages, as well as groups speaking languages belonging to the Tibeto-Chinese language family (due to Mongolian influence in Assam) and Austro-Asiatic (Munda) speakers. A separate cluster contained two groups of tribals, speaking Munda and Bodo (Tibeto-Chinese) respectively. Language replacement is a possibility that exists as a potential cause of this discrepancy (Walter et al., 1986).

Relationship of Indians to Other World Populations:

Regarding the presence of Australoid type of people in southern India, the theory has been proposed that these are the vestiges of the movement of people from Africa to Australia, by way of southern India. It has been shown that people of Dravidian descent are not genetically related to Australians (Cavalli-Sforza et al., 1994; Kirk and Thorne, 1976).

Indians have been suggested to be genetically more closely related to Mongoloids than Caucasoids or Negroids, based on protein and enzyme allele frequency data (Roychoudhury, 1977). Another gene frequency study placed the
Indian population closer to Mongolians, Africans, Afghans than to Caucasians (Sahai et al., 1978).

A brief summary of the results obtained in the studies described in this section is given in table 2.

**MtDNA Sequence Analysis in Human Population Studies**

There are a number of advantages to the use of sequence analysis of the mitochondrial D-loop in human populations studies. Human mitochondrial DNA (mtDNA) consists of a circular genome 16 569 bp in length, which displays strict maternal inheritance (Giles et al., 1980). Since there is no recombination within the mitochondrial genome (DeFrancesco et al., 1980), mutation remains the sole source of mtDNA sequence variation (Aquadro and Greenberg, 1983). This provides an opportunity for the examination of evolutionary relationships without having to address the confounding effects of recombination.

The rate of evolution of the animal mitochondrial genome is approximately 10 times greater than that of the nuclear genome (Brown et al., 1979). This rapid rate of evolution leads to a sufficient amount of differentiation between closely related populations.

Of particular interest to human evolutionary studies is the major non-coding region of mtDNA. This 1122 bp region of the mitochondrial genome, known as the control region, lies between the tRNA\(^{\text{Pro}}\) and tRNA\(^{\text{Phe}}\) genes. It contains the Origin of Replication, which marks the beginning of the displacement loop, or D-loop.
(Anderson et al., 1981). Sequence variability within this region is higher than for the remainder of the mitochondrial genome (Vigilant et al., 1989), and has been demonstrated to be quite unequal over the region (Horai and Hayasaka, 1990). These properties of the mitochondrial D-loop region make it a useful tool in the study of relationships between even closely related individuals. These factors, along with the advent of the Polymerase Chain Reaction (Saiki et al., 1988), which has simplified the obtainment of large numbers of sequences of reasonable length, have led to a proliferation of mitochondrial DNA studies on various human populations in an attempt to reconstruct historical and anthropological events.

Pairwise Comparisons of Mitochondrial DNA Sequences as an Indicator of Population History

The distribution of pairwise differences (or mismatch distribution) for a population of roughly constant size tends to be bi- or multi-modal. This is because after the first branching event, approximately half of the samples are on opposite sides of the common root. As a result of this, pairwise differences between sequences from the same side of the root will be substantially smaller than pairwise differences between sequences from opposite sides of the root. The result of this would be a bimodal distribution of differences (Slatkin and Hudson, 1991).

A population that has grown exponentially will exhibit a roughly Poisson distribution of pairwise mismatch frequencies (Marjoram and Donnelly, 1994; Slatkin and Hudson, 1991). This occurs because exponential growth forces coalescent events into a narrow time period. The concurrent event to this is the obtainment of a star-
like phylogeny.

One cannot however conclude from a Poisson distribution that there has been exponential growth, as the distribution could be the result of any event that has forced most coalescent events into a narrow time period (Slatkin and Hudson, 1994). Possible factors other than growth rate affecting mismatch distributions include the effects of migration and the original population size.

Under an island model of migration with four colonies, reasonable amounts of migration (such as a migration probability of 0.01) in an exponentially growing population, are enough to allow the population to mimic a randomly mating population, leading to a roughly Poisson distribution. Under the same model with only small rates of migration (such as a migration probability of 0.001), the exponentially growing population will however exhibit a multimodal distribution. Very minute rates of migration again yield Poisson distributions (Marjoram and Donnelly, 1994).

A population that was very small originally, and was growing at a constant rate followed by a burst of exponential growth, will exhibit unimodal distributions approaching Poisson, but if the population size is quite large prior to exponential growth, distributions will still be multimodal, despite the exponential growth spurt.

Bottlenecks and founder effects, as well as natural selection will also lead to unimodal distributions (Marjoram and Donnelly, 1994; Slatkin and Hudson, 1991).
Previous Human Population Studies Based on mtDNA

The analysis of mtDNA sequences from various human populations has traditionally been used for the following purposes:

1. *Establishing relationships between human populations*

   Trees can be inferred for various human populations, in order to attempt to elucidate genetic relationships between populations. A number of studies have addressed this question with regard to various human populations. Generally, trees have few clusters which are specific to only one population (DiRienzo and Wilson, 1991). The highest degree of differentiation is seen in African populations, which tend to show a more clear separation from the remainder of human populations. Native Americans also tend to form more or less separate clusters, but the clusters are distributed among the remainder of the world populations. Caucasian populations generally exhibit starlike clusters, where most branches originate at a common centre, possibly indicating rapid growth at some point in history (Cann *et al.*, 1987; DiRienzo and Wilson, 1991; Horai *et al.*, 1991; Vigilant *et al.*, 1989)

2. *Determination of the age and geographical origin of modern H.sapiens*

   The mitochondrial Eve theory, suggesting that all human beings can trace their ancestry to a single female living in Africa 200 000 years ago has received considerable attention. The theory was proposed on the basis of mtDNA sequence data, which showed African samples as having deeper branches as well as larger sequence differences, thus suggesting greater
antiquity. The ancestral mtDNA to which all humans can trace back is therefore assumed to be of African origin. An assumption of a constant mutation rate allowed the dating of this ancestor to approximately 200,000 years ago (Cann et al., 1987; Vigilant et al., 1991).

3. Reconstruction of previous population expansion events

Simulations of pairwise sequence difference distributions under various models of population growth suggest that population demography histories can be inferred by looking at these distributions (Marjoram and Donnelly, 1994; Slatkin and Hudson, 1991). Studies using empirical data reveal multimodal distributions for samples of African, Native American and Japanese descent, suggesting that these populations have maintained a relatively constant size over time (DiRienzo and Wilson, 1991; Horai et al., 1993). Only Caucasian populations exhibited a unimodal, Poisson type distribution, possibly suggesting a previous expansion event (Horai et al., 1993).

4. Resolving debates involving the origins of various human populations

The origins of populations such as Native Americans have been postulated on the basis of archeological evidence. MtDNA analysis can be used to attempt to resolve these debates on the basis of genetic heritage. Findings are able to suggest that the settlement of the Americas occurred in four waves of migrations, dated to between 14,000 and 21,000 years ago, which is in agreement with archeological evidence (Horai et al., 1993).

Few mtDNA D-loop studies have addressed populations as closely related
as the Indian population. Some mtDNA sequence work has recently been done on south Indian samples (Mountain et al., 1995). Phylogenetic analysis of these sequences (from Karnataka and Kerala) compared to those of other populations reveals that south Indian sequences cluster with other samples of non-African origin, in a starlike cluster (Mountain et al., 1995), as do other samples of Caucasian origin.

A 9 bp deletion of the intergenic region between the COII and tRNA genes, which has previously been reported to be prevalent in Asians, including Chinese, Japanese and Native Americans (Horai et al., 1993), was found in 27% of Asians, 8% of South Indians, and 0% of North Indians, Bangladeshis and Pakistanis (Melton et al., 1995). This may indicate an influx at some point of Asian genes into South India, but not into North India.

Objectives of the Present Study

The objectives of the present study are to attempt to resolve any genetic differences between individuals of different geographical origins, castes and linguistic groups. Although gene frequency studies such as those previously performed can provide simple analyses of interrelationships between different populations, data from different genes do not always give the same results. Some markers may also be under selection. Previous studies have also been mostly regionally based, with few addressing the whole of the subcontinent. In this study we use a different approach, the sequencing of a non-coding, and therefore
hopefully selectively neutral, region of mitochondrial DNA. We also address the entire Indian subcontinent, rather than just one region.

The majority of previous studies involving the use of mtDNA sequencing of the hypervariable D-loop to reconstruct population histories for populations have generally dealt with populations that are further diverged than the Indian population. In this study, we attempt to determine if sequencing of the hypervariable domain of the mitochondrial D-loop is an analytical tool sensitive enough to resolve differences between more closely related individuals, such as those found within the Indian subcontinent.

Should this technique prove to be sensitive enough to detect differences between populations this closely related, we would like to be able to determine a little more about the sources of the grand amount of diversity observed in the Indian population. To this end we address geographical origin, language spoken and caste in all of our analyses as potential factors involved in either the creation or the maintenance of genetic diversity.

Additionally, we would like to determine if the sensitivity is sufficient to enable us to pick up on any population substructuring of the extent which may be seen as a result of a social stratification system such as the Hindu caste system. It would be of interest to determine if the genetic differences are greater between castes than between geographical regions within the same caste. Lack of resolution with respect to geographical origin may exist if members of castes are more closely related to members of the same caste from different geographical
origins than to members of different castes from their proximal geographical region. This could occur if the intercaste marriage restrictions have been as stringent as they supposed to be, in which case it would be preferable to marry an individual from outside the region so long as they are still within the same caste, than to marry outside of the caste within the same region.
Materials and Methods

Sampling procedure

Blood samples were obtained on a volunteer basis from individuals of Indian origin residing in Canada or in India. All donors filled out questionnaires to determine their place of birth, native language and religion. These criteria were determined for the individual providing the sample, as well as for both of their parents.

The geographical origins of the 84 individuals sequenced in this study are shown on the map in figure 3. Distribution of samples on the basis of mother's province of birth, mother's language and religion are shown in tables 3-5.

Blood samples were collected in 5 ml VACUTAINER brand blood collection tubes containing 0.05 ml of 15% EDTA (K$_3$) solution. Blood was stored in 1 ml aliquots at -70°C.

DNA Extraction

Genomic DNA was extracted from 1 ml of whole blood as follows: Red blood cells were lysed 3 times using an NH$_4$Cl / Na$_2$HCO$_3$ lysis buffer and the white blood pellet obtained was rinsed with saline solution. The white blood pellet was lysed using SSTE and Proteinase K (Boehringer-Mannheim). Incubation for
Figure 3: Geographic distribution of samples.

Geographical locations of the 84 samples sequenced in this study. Sample numbers in brackets represent individuals for whom exact geographical origin is not known beyond the mother's province of birth.
Sample Distribution
Table 3: Geographical distribution of individuals sequenced in this study.

<table>
<thead>
<tr>
<th>Province</th>
<th># of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Southern</strong></td>
<td>13</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>1</td>
</tr>
<tr>
<td>Tamil Nadu</td>
<td>6</td>
</tr>
<tr>
<td>Kerala</td>
<td>4</td>
</tr>
<tr>
<td>Karnataka</td>
<td>1</td>
</tr>
<tr>
<td>Andhra Pradesh</td>
<td>1</td>
</tr>
<tr>
<td><strong>Western</strong></td>
<td>13</td>
</tr>
<tr>
<td>Goa</td>
<td>1</td>
</tr>
<tr>
<td>Maharashtra</td>
<td>4</td>
</tr>
<tr>
<td>Madhya Pradesh</td>
<td>1</td>
</tr>
<tr>
<td>Gujarat</td>
<td>5</td>
</tr>
<tr>
<td>Pakistan</td>
<td>4</td>
</tr>
<tr>
<td><strong>Eastern</strong></td>
<td>15</td>
</tr>
<tr>
<td>West Bengal</td>
<td>13</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>1</td>
</tr>
<tr>
<td>Assam</td>
<td>1</td>
</tr>
<tr>
<td><strong>Northern</strong></td>
<td>43</td>
</tr>
<tr>
<td>Bihar</td>
<td>3</td>
</tr>
<tr>
<td>Uttar Pradesh</td>
<td>30</td>
</tr>
<tr>
<td>Haryana</td>
<td>2</td>
</tr>
<tr>
<td>Punjab</td>
<td>3</td>
</tr>
<tr>
<td>Himachal Pradesh</td>
<td>1</td>
</tr>
<tr>
<td>Kashmir</td>
<td>1</td>
</tr>
<tr>
<td>Nepal</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>84</td>
</tr>
</tbody>
</table>
Table 4: Distribution of languages spoken by individuals sequenced in this study, separated on the basis of belonging to either the Dravidian or Indo-Iranian language families.

<table>
<thead>
<tr>
<th>Language</th>
<th># of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dravidian</td>
<td>13</td>
</tr>
<tr>
<td>Indo-Iranian</td>
<td>65</td>
</tr>
<tr>
<td>Unknown</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total known</strong></td>
<td><strong>78</strong></td>
</tr>
</tbody>
</table>

Table 5: Distribution of religions and castes for individuals sequenced in this study.

<table>
<thead>
<tr>
<th>Religion / Caste</th>
<th># of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hindu</strong></td>
<td>58</td>
</tr>
<tr>
<td>Brahmin</td>
<td>29</td>
</tr>
<tr>
<td>Kshatriya</td>
<td>13</td>
</tr>
<tr>
<td>Vaisya</td>
<td>16</td>
</tr>
<tr>
<td>Scheduled Caste</td>
<td>0</td>
</tr>
<tr>
<td><strong>Muslim</strong></td>
<td>6</td>
</tr>
<tr>
<td><strong>Christian</strong></td>
<td>7</td>
</tr>
<tr>
<td><strong>Unknown</strong></td>
<td>13</td>
</tr>
<tr>
<td><strong>Total known</strong></td>
<td><strong>71</strong></td>
</tr>
</tbody>
</table>
1 hour at 55°C was followed by phenol:chloroform extraction, consisting of two extractions using equal volumes of equilibrated phenol, one extraction using a 1:1 mixture of phenol:chloroform iso-amyl alcohol, and one extraction using only chloroform iso-amyl alcohol. DNA was precipitated with 2-3 volumes of 100% Ethanol and 1/10th volume of 2M NaAc. The DNA pellets were rinsed with 70% ethanol, dried and dissolved in 50 μl of TE. Samples are stored at -20°C.

**PCR Amplification of 1024 bp Region of Mitochondrial Genome**

A 1024 bp fragment of the mitochondrial genome containing the non-coding D-loop region was amplified using the Polymerase Chain Reaction. The region amplified lies between base pairs 15975 - 429, numbered according to the human reference sequence (Anderson et al., 1981). Figure 4 gives the complete reference sequence of the region amplified. The fragment was amplified using the primers L15996 5'TCCACCATTAGCACCCAAAGC 3' and H408 5'CTGTTAAAAGTGCATACCGCCA 3', where L and H refer to Light and Heavy strands and the number refers to the 3' end of the primer. All primers were synthesized at the Mobix Central Facility at McMaster University. Primers were diluted in distilled deionized water to a 10 μM concentration.

PCR amplifications were carried out in thin-walled GeneAmp™ reaction tubes, using the GeneAmp® PCR Reagent Kit, containing AmpliTaq® DNA Polymerase (all from Perkin Elmer). Reactions were carried out in 50 μl volumes following the manufacturer's instructions with some minor modifications. DNA
Figure 4: Reference sequence of mitochondrial D-loop.

Reference sequence* of 1024 bp region of mitochondrial D-loop. Primers used to amplify the complete 1024 bp region as well as the internal primers used to sequence the 195 bp region are shown in bold. The start and end of the 195 bp region sequenced in this study are indicated by bold capitalized letters.

* Anderson et al., 1981.
ttttatacag
samples were preheated for 5 minutes at 95°C in order to denature the DNA. Reactions contained 0.25 μl (1.25 Units) of AmpliTaq® DNA Polymerase, 34.75 μl of distilled deionized water, 5 μl of the supplied Reaction Buffer (containing 100 mM Tris-HCl, pH 8.3 @ 25°C and 500 mM KCl), 3 μl MgCl₂ solution (25 mM MgCl₂), 1 μl of each dNTP for a final concentration of 200 μM, 1 μl of each primer, for a final concentration of 0.5 μM, and 1 μl of template DNA (approximately 100 ng). Reactions were overlaid with 25 μl mineral oil and briefly centrifuged to collect liquid at bottoms of tubes. The cycling reactions were performed using the Perkin Elmer Cetus DNA Thermal Cycler 480. A negative control was included in all amplification procedures. Reactions were preheated for 1 minute at 95°C, followed by 30 denature-anneal-extend cycles of 45 seconds at 93°C, 1 minute at 55°C, and 3 minutes at 74°C.

PCR products were visualized by running 5 μl of each reaction on 1.2% agarose minigels stained with ethidium bromide. DNA fragments were sized in comparison to a 100-bp ladder (Gibco-BRL). Products which displayed a single band of the correct size were purified using the GlassMAX™ DNA Isolation Spin Cartridge System (Gibco-BRL), and the purified DNA was resuspended in 40 μl of TE.

### Sequencing of 195 bp Region of Mitochondrial D-loop

The region sequenced was a 195 bp hypervariable region located between basepairs 16125 - 16319. The primers used were
Primer naming conventions are the same as above.

The purified PCR products were cycle sequenced using the AmpliCycle™ Sequencing Kit (supplied by Perkin Elmer), following the manufacturer's instructions. Primers were radioactively end-labelled with $^{33}$P, by combining in a thin-walled GeneAmp™ reaction tube; 3 µl of T4 Polynucleotide Kinase (10 Units/µl), 1 µl of the appropriate primer, 0.6 µl of 10X Kinase Buffer, 0.6 µl of distilled deionized water, and 1 µl of $\gamma$-$^{33}$P-ATP. Tubes were incubated at 37°C for 15 minutes followed by 90°C for 5 minutes. Labelled primers were stored at -20°C for up to 5 days.

For every sample to be sequenced, termination reaction mixes were prepared in thin-walled PCR tubes for each of the 4 nucleotides. Termination mixes contained 2 µl of the 7-deaza-dNTP, and 6 µl of the appropriate reaction mix. Reaction mixes consisted of 4 µl of 10X Cycling Mix (containing 1 Unit of AmpliTaq® DNA polymerase), 20 µl of distilled deionized water, 5 µl of template DNA (10-100 fmoles) and 1 µl of the $^{33}$P-end-labelled primer (1.6 pmoles/µl). Reactions were mixed and overlaid with 15 µl mineral oil.

Sequencing was carried out using the Perkin-Elmer Cetus DNA Thermal Cycler 480, with the following extension programs: One cycle of 1 minute at 95°C, followed by 20 denature-extend cycles of 1 minute at 95°C and 1 minute at 55°C for primer L16096, and 20 cycles of 1 minute at 95°C and 1 minute at 65°C for primer H16379. After amplification, 4 µl of Stop Solution (95% Formamide; 20 mM
EDTA; 0.05% Bromophenol Blue; 0.02% Xylene Cyanole FF) were added to each reaction tube, and the tubes were briefly centrifuged to mix and collect the liquid at the bottoms of the tubes. Reactions were stored for up to 1 week at -20°C.

Sequencing reaction products were run on 8% polyacrylamide gels containing 7 M Urea, on the Life Technologies Inc. Model S2 sequencing apparatus, at 2000 Volts, 65 Watts and 65 mAmps. Gels were run for approximately 1.75 hours for short runs and 3.5 hours for long runs. These two runs were sufficient to clearly resolve all 195 bp of the target sequence. Gels were dried on the BioRad Model 583 gel dryer for 1.25 hours, and autoradiographed using Kodak Scientific X-OMAT AR X-ray film. Figure 5 shows an example of a sequence obtained using the above procedure.

**Sequence Analyses**

Virtual heterozygosity, or haplotype diversity was computed using the following formula (Hartl and Clark, 1989):

\[ 1 - \sum p_i^2 \]

where \( p_i \) is the frequency of the \( i \)-th haplotype.

**Phylogenetic Analysis**

For the phylogenetic analyses, three major factors were addressed, these being geographical origin, language and caste. All 84 sequences were included in the geographical analysis. For language, 78 sequences were included (only
Figure 5: Image of autoradiograph of sequence.

Image of a sequence obtained using cycle-sequencing (lane order is A C G T). Reaction products were run on an 8% acrylamide gel and visualized by autoradiography. The regions corresponding to the 195 bp region analyzed in this study lie between the arrows.
short run

long run
those for which language was known), and for caste analysis, 71 sequences were included (again only those for whom caste was known). Geographical and caste analyses were repeated including 78 published sequences from Karnataka and Kerala (Mountain et al., 1995). Note that only those published sequences for which all 195 bp were resolved were included in any analyses.

Additional sequences from individuals of European, Middle Eastern, African and miscellaneous Caucasian descent were obtained via Genbank and EMBL. See table 6 for the sources of these sequences. These were used in order to generate a global human phylogeny including Indian sequences.

Sequences were aligned using CLUSTAL V (Higgins and Sharp, 1988) and by hand. For the phylogenetic analyses the PHYLIP software package was used (Felsenstein, 1993). Genetic distance was computed using DNADIST with the following parameters: Distances were estimated by Kimura's two-parameter model, assuming a transition:transversion ratio of 25:1 for the 84 sequences obtained in this study, a 20:1 ratio for the analyses including the 78 published sequences from south India, and a 16:1 ratio for the global analysis including the remaining published sequences. Trees were inferred using the Neighbor-Joining method (Saitou and Nei, 1987) in the program NEIGHBOR and the Maximum Parsimony method in DNAPARS. Comparison of trees using Templeton's Delta Q test was performed using the User Tree function in DNAPARS.

Bootstrap analysis was carried out using SEQBOOT, and the samples were run through NEIGHBOR or DNAPARS to obtain trees. The consensus trees were
Table 6: Origins of sequences used in phylogenies.

<table>
<thead>
<tr>
<th>Origin</th>
<th># Sequences</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pygmy</td>
<td>14</td>
<td>Vigilant <em>et al.</em>, 1991</td>
</tr>
<tr>
<td>South Indian</td>
<td>78</td>
<td>Mountain <em>et al.</em>, 1995</td>
</tr>
<tr>
<td>Indian (various origins)</td>
<td>84</td>
<td>Present study</td>
</tr>
<tr>
<td>European (northern)</td>
<td>1</td>
<td>Present study</td>
</tr>
<tr>
<td>European</td>
<td>1</td>
<td>Anderson <em>et al.</em>, 1980</td>
</tr>
<tr>
<td>European (various)</td>
<td>14</td>
<td>Vigilant <em>et al.</em>, 1991</td>
</tr>
<tr>
<td>European (Sardinian)</td>
<td>45</td>
<td>DiRienzo <em>et al.</em>, 1991</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>35</td>
<td>DiRienzo <em>et al.</em>, 1991</td>
</tr>
</tbody>
</table>
determined using CONSENSE. Trees were rearranged using RETREE, and drawn using DRAWGRAM. Trees with symbols were drawn using TREETOOL 1.0 (Maciukenas, 1991), and the resulting postscript files were edited to include symbols and cluster labels.

In each of the phylogenetic analyses, 14 sequences from West and East Pygmies (Vigilant et al., 1991) were included to root the samples. Pygmies were chosen as they cluster away from Europeans and Indians (Vigilant et al., 1989), yet are not too far removed genetically to cause the resolution of differences between Indian individuals to be lost.

For the sake of geographic comparisons, individuals were grouped arbitrarily as being of northern, western, eastern or southern origin. For the geographic locations of these groups see figure 6.

**Distributions of Frequencies of Pairwise Differences**

The number of pairwise differences between all sequence pairs were computed for various subgroups of the population under study. Mismatch distributions were determined for all 84 samples, and for Brahmins, Kshatriyas, Vaisyas, Muslims and Christians. The frequencies of the number of pairwise differences were computed using the algorithm given in Appendix C, and normalized by setting the area of each curve equal to 1. These adjusted frequencies were plotted as a function of number of differences.

Sample means (\( \bar{x} \)) and sample variances (\( s^2 \)) were computed for each
Figure 6: Geographical regions of India.

Geographic subdivision of India on the broad basis of northern, western, eastern, southern and central segments. Definitions are mostly arbitrary except for the southern segment, which is based on Dravidian descent. Due to the small number of samples from the central provinces, these samples have been grouped with the western samples for the purpose of the present study.
Geographical Regions

- Northern
- Western
- Eastern
- Southern
- Central
distribution. The coefficient of dispersion for each subgroup was computed as well, where $C.D. = \frac{s^2}{\bar{x}}$. The coefficient of dispersion provides a rough estimate of the fit of a sample distribution to a Poisson curve. As the mean is equal to the variance under the Poisson model, values for C.D. of approximately 1 suggest a reasonable fit to a Poisson distribution.

Statistical deviation of the observed distribution from one predicted under the Poisson model for the computed sample mean was determined for the whole population, Brahmins, Kshatriyas and Vaisyas using a Chi-square Goodness-of-fit test. For classes of numbers of pairwise differences where the expected values were less than 5, classes were lumped together to give a minimum expected value of > 5. Degrees of freedom were computed as equalling the number of classes - 2 (1 degree of freedom being lost due to $\mu$ having been estimated from the sample mean). Chi-square analysis was not performed on the Muslim and Christian subgroups as these groups contained less than 30 pairwise comparisons.

**Polymorphic Sites**

Polymorphic sites were defined as sites where the frequency of the most common nucleotide was < 95%. The occurrences of substitutions were plotted on maps of India. Chi-square analyses were performed to determine if there was any significant difference in frequency of substitutions with respect to these sites between northern and southern samples. Southern samples were defined as those located in the provinces Kerala, Tamil Nadu, Karnataka, Andhra Pradesh
and in Sri Lanka, or, alternatively, as those provinces of whom the inhabitants are speakers of the Dravidian languages.

See Appendix for detailed protocols of the procedures described in this section.
Results

General Summary of Sequence Analysis

Sequences of a 195 bp hypervariable region of the mitochondrial D-loop were obtained for 84 individuals of Indian origin. Out of 195 sites, 54 were variable, corresponding to 28% of the sites. Transitions alone were found at 50 of these sites and transversions alone at 2 sites. One site (site 16265) had both transitions and transversions. There was one individual with a deletion at site 16166. See table 7 for a summary of sequence variation within the individuals sequenced in this study.

Out of 84 individuals, 61 unique types of sequences were found. Figure 7 lists all of the sequences. The table is presented in the order given by the phylogenetic tree shown in figure 8.

Virtual heterozygosity, or haplotype diversity (Hartl and Clark, 1989) was calculated to be 0.96 for the whole population. Tables 9 & 10 list the heterozygosities computed for the different castes and religious groups investigated in this study, as well as for northern vs. southern Indians. Heterozygosity for southern Indians was lower ($h = 0.90$) than for northern Indians ($h = 0.97$).
Table 7: Sequence variability within the Indian population for the 195 bp mtDNA region sequenced in this study.

<table>
<thead>
<tr>
<th>Sample Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>84</td>
</tr>
<tr>
<td>Variable Sites</td>
<td>54</td>
</tr>
<tr>
<td>Sites with Transitions</td>
<td>50</td>
</tr>
<tr>
<td>$A \leftrightarrow G$</td>
<td>18</td>
</tr>
<tr>
<td>$T \leftrightarrow C$</td>
<td>33</td>
</tr>
<tr>
<td>Sites with Transversions</td>
<td>2</td>
</tr>
<tr>
<td>$A \rightarrow T$</td>
<td>$1^a$</td>
</tr>
<tr>
<td>$A \rightarrow C$</td>
<td>1</td>
</tr>
<tr>
<td>Sites with both Transitions and Transversions</td>
<td></td>
</tr>
<tr>
<td>$A \rightarrow G$ &amp; $A \rightarrow C$</td>
<td>1</td>
</tr>
<tr>
<td>Sites with Deletions</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ At this site 8/9 individuals have an $A \rightarrow T$ transversion, and 1/9 individuals has an $A \rightarrow C$ transversion.
Figure 7: Variable sites for 84 mitochondrial D-loop sequences.

Variable sites for the 195 bp sequences of the 84 individuals sequenced in this study. Sequences are arranged according to the UPGMA tree shown in figure 8. The numbers at the top of the chart represent the last 3 digits of the numbering system employed for Anderson's reference sequence (Anderson et al., 1981). Sequences are aligned with respect to Anderson's reference sequences (000 And). The numbers at the bottom of the chart represent the number of individuals with substitutions at each site. Asterices correspond to sites with either transversions or both transitions and transversions. The name codes are interpreted as follows:

eg. 001 W-B V
     ↑   ↑   ↑
     sample province caste
     id. of origin

Table 8 lists the definitions of the abbreviations used in name codes.
Sites with transversions or both transversions and transitions.
Figure 8: UPGMA tree of all 84 sequences.

Phylogenetic tree used to arrange sequences shown in figure 6. This tree was inferred using the UPGMA algorithm employing genetic distances computed according to Kimura's two-parameter model (Kimura, 1980).
Table 8: Definitions of sample code abbreviations.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-P</td>
<td>Andhra Pradesh</td>
</tr>
<tr>
<td>Asm</td>
<td>Assam</td>
</tr>
<tr>
<td>Bih</td>
<td>Bihar</td>
</tr>
<tr>
<td>E-B</td>
<td>Bangladesh</td>
</tr>
<tr>
<td>Goa</td>
<td>Goa</td>
</tr>
<tr>
<td>Guj</td>
<td>Gujarat</td>
</tr>
<tr>
<td>H-P</td>
<td>Himachal Pradesh</td>
</tr>
<tr>
<td>Har</td>
<td>Haryana</td>
</tr>
<tr>
<td>Kar</td>
<td>Karnataka</td>
</tr>
<tr>
<td>Kas</td>
<td>Kashmir</td>
</tr>
<tr>
<td>Ker</td>
<td>Kerala</td>
</tr>
<tr>
<td>M-P</td>
<td>Madhya Pradesh</td>
</tr>
<tr>
<td>Mah</td>
<td>Maharashtra</td>
</tr>
<tr>
<td>Nep</td>
<td>Nepal</td>
</tr>
<tr>
<td>Pak</td>
<td>Pakistan</td>
</tr>
<tr>
<td>Pun</td>
<td>Punjab</td>
</tr>
<tr>
<td>Sri</td>
<td>Sri Lanka</td>
</tr>
<tr>
<td>T-N</td>
<td>Tamil Nadu</td>
</tr>
<tr>
<td>U-P</td>
<td>Uttar Pradesh</td>
</tr>
<tr>
<td>W-B</td>
<td>West Bengal</td>
</tr>
</tbody>
</table>

B Brahmin
K Kshatriya
V Vaisya
m Muslim
c Christian
- / ? Unknown
SC Scheduled Caste\textsuperscript{a}
Tr Tribal\textsuperscript{a}

\textsuperscript{a} Sequences from Mountain et al., 1995.
Table 9: Heterozygosities for different castes and religions within the 84 samples analysed in this study, based on Hartl and Clark’s formula for haplotype diversity (Hartl and Clark, 1989).

<table>
<thead>
<tr>
<th></th>
<th>Brahmin</th>
<th>Kshatriya</th>
<th>Vaisya</th>
<th>Muslim</th>
<th>Christian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Individuals</td>
<td>29</td>
<td>13</td>
<td>16</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Heterozygosity</td>
<td>0.96</td>
<td>0.90</td>
<td>0.91</td>
<td>0.83</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 10: Heterozygosities for northern vs. southern Indians.

<table>
<thead>
<tr>
<th></th>
<th>Northern Indians</th>
<th>Southern Indians</th>
<th>Total Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Individuals</td>
<td>71</td>
<td>13</td>
<td>84</td>
</tr>
<tr>
<td>Heterozygosity</td>
<td>0.97</td>
<td>0.90</td>
<td>0.96</td>
</tr>
</tbody>
</table>
**Phylogenetic Analysis**

Phylogenetic trees were inferred from the Indian sequences obtained in this study for the following criteria: Geographical origin, language and caste. Additional trees incorporating 78 sequences of south Indian origin from Mountain et al.'s (1995) study were inferred for trees addressing geographical origin and caste. All trees include 14 sequences from Eastern and Western Pygmies (Vigilant et al., 1991). In all cases, the Pygmies formed an entirely separate and distinct cluster from the Indian subjects.

Trees were inferred using the Neighbor-Joining algorithm (Saitou and Nei, 1987) as well as the Maximum Parsimony method. In all cases but one, Templeton's Delta Q test showed no significant difference between trees inferred using the Neighbor-Joining method and those inferred using the Maximum Parsimony method. In most cases, the difference in number of steps required to generate the trees is 0 or 1 step. Detailed results of the Delta Q tests of significance are given in table 11.

Some branches in all trees are negative in length. In all cases, the negative values were small, and branches are assigned a value of zero for length (Saitou and Nei, 1987). Some degree of clustering is seen in all trees, and the larger clusters are defined by visual inspection and marked on the phylogenies for reference.

**Geographical Origin**

A phylogenetic tree was inferred for all 84 samples on the basis of
Table 11: Steps required for different tree-inference methods. Significances were computed using Templeton's Delta Q test, in the PHYLIP Phylogeny inference program (Felsenstein, 1993).

<table>
<thead>
<tr>
<th>Population</th>
<th>Tree</th>
<th>Steps</th>
<th>Diff. Steps</th>
<th>S.D.</th>
<th>Sig. Diff.?</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study Geography</td>
<td>Neighbor-Joining</td>
<td>124</td>
<td>1</td>
<td>2.6526</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Maximum Parsimony</td>
<td>123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>Neighbor-Joining</td>
<td>122</td>
<td>1</td>
<td>3.3252</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Maximum Parsimony</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caste</td>
<td>Neighbor-Joining</td>
<td>112</td>
<td>1</td>
<td>3.0077</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Maximum Parsimony</td>
<td>111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian* Geography</td>
<td>Neighbor-Joining</td>
<td>153</td>
<td>0</td>
<td>3.7513</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Maximum Parsimony</td>
<td>153</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caste</td>
<td>Neighbor-Joining</td>
<td>142</td>
<td>31</td>
<td>7.0162</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Maximum Parsimony</td>
<td>111</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indian + other populations</td>
<td>Neighbor-Joining</td>
<td>245</td>
<td>0</td>
<td>5.6714</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Maximum Parsimony</td>
<td>245</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Individuals from this study plus published sequences from southern Indians (Mountain et al., 1995).
geographical origin. Two versions of this phylogeny are given in figures 9 & 10. Terminal nodes on figure 9 are labelled with information with respect to both geographic origin and caste. Two major clusters are seen (clusters A and B). Individuals of all geographic origins are represented within each cluster. Chi-square analysis of clusters comparing expected numbers of representatives of each geographic origin vs. the numbers observed reveals no significant difference between expectation and observation.

The tree inferred from the 162 Indian sequences (see figure 11) includes individuals of northern and of southern Indian origin in more equal proportions (71 northern Indians and 91 southern Indians). Again, two major clusters are observed. Each cluster, except for the small cluster (cluster D) seen near the top of the phylogeny, includes individuals representing all geographic origins. Cluster D, containing 7 individuals, includes only individuals of southern descent. These 7 individuals represent various castes and religions (1 Brahmin, 3 Scheduled Caste, 1 Christian and 2 Tribal). Chi-square analysis reveals no significance with respect to distribution of individuals of various geographic origin within clusters.

Language

Language was known for 78 out of the 84 subjects. The phylogeny inferred from these 78 sequences is shown in figure 12. Individuals were grouped according to whether they speak a Dravidian or an Indo-Iranian language. Three major clusters are observed (A, B and C), as well as two small clusters containing
Figures 9 - 16: Phylogenetic trees inferred using the Neighbor-Joining algorithm (Saitou and Nei, 1987). The genetic distance matrices required by this algorithm were computed using Kimura's two-parameter model (Kimura, 1980). All trees were generated by the software package PHYLIP. Trees with symbols were plotted with the software package TREETOOL.a

* See materials and methods for a detailed description of programs used.

Figure 9: Phylogenetic tree - 84 Indian individuals.

Phylogenetic tree inferred from 84 Indians sequenced in this study, plus 14 Pygmies (Vigilant et al., 1991).
**Figure 10:** Phylogenetic tree - 84 Indian individuals, based on geography.

Phylogenetic tree identical to the tree shown in *figure 9*. Symbols correspond to the following subgroups:

- $\Delta$  Pygmies
- $\bullet$  Northern Indians$^a$
- $\circ$  Western Indians$^a$
- $\Diamond$  Eastern Indians$^a$
- $\ast$  Southern Indians$^a$

$^a$ According to figure 6.
**Figure 11:** Phylogenetic tree - 162 Indian individuals, based on geography.

Phylogenetic tree includes 84 Indians sequenced in this study, 78 south Indians from Mountain *et al.* (1995), and 14 Pygmies (Vigilant *et al.*, 1991). Symbols correspond to the following subgroups:

- Δ Pygmies
- ♦ Northern Indians
- ○ Western Indians
- ◊ Eastern Indians
- • Southern Indians

* According to figure 6.
Figure 12: Phylogenetic tree - 78 Indian individuals, based on language.

Phylogenetic tree inferred for the 78 Indians sequenced in this study for whom language was known, plus 14 Pygmies (Vigilant et al., 1991). Symbols correspond to the following subgroups:

△ Pygmies
♦ Indo-Iranian language speakers
○ Dravidian language speakers
1 and 2 individuals respectively. All major clusters include speakers of both language families. However, the representation of Dravidian language speakers in cluster B is lower than expected. This discrepancy is significant at the $p < .05$ level according to Chi-square analysis ($\chi^2 = 4.47$, d.f. = 1).

Caste

Caste or religion was known for 71 out of 84 individuals. The phylogeny inferred from these sequences is shown in figure 13. Individuals belonged to either one of 3 major Hindu castes (29 Brahmins, 13 Kshatriyas and 16 Vaisyas), or were Muslim (6 individuals) or Christian (7 individuals). This tree has one major cluster (B) and 2 smaller clusters (A and C). Again, clusters tend to include individuals of various castes and religions. However, cluster A is of interest in that out of 14 individuals, there are no members of the Vaisya caste within the cluster. Chi-square analysis of the representation of all 5 religious categories reveals no significant deviation from expectation. A Chi-square test of the expected proportion of Vaisyas in cluster A vs. all other religions was significant at the $p < .05$ level ($\chi^2 = 4.07$, d.f. = 1). The mean frequency of pairwise differences between sequence pairs within this group however, is not lower for this caste than for the other caste or religious groups (see table 12).

Inclusion of the 78 southern Indian samples from Mountain et al.'s (1995) results in the phylogenetic tree shown in figure 14. Two extra categories are included in this analysis, the Scheduled Caste, and individuals of Tribal origin. There appears to be no clustering on the basis of caste or religion within this
Figure 13: Phylogenetic tree - 71 Indian individuals, based on caste.

Phylogenetic tree inferred for 71 Indians sequenced in this study, plus 14 Pygmies (Vigilant et al., 1991). Symbols correspond to the following subgroups:

- △ Pygmies
- ♦ Brahmin
- ♥ Kshatriya
- o Vaisya
- ◇ Muslim
- x Christian
Figure 14: Phylogenetic tree - 149 Indian individuals, based on caste.

Phylogenetic tree includes 71 of the Indians sequenced in this study, 78 south Indians from Mountain et al. (1995), and 14 Pygmies (Vigilant et al., 1991). Symbols correspond to the following subgroups:

- Δ Pygmies
- ♦ Brahmin
- ♥ Kshatriya
- ○ Vaisya
- ▽ Scheduled Caste
- τ Tribal
- ◊ Muslim
- x Christian
phylogeny.

Caste analysis on northern Indians only

The caste analysis was repeated using only sequences from individuals from northern India, as the caste system has been in existence for longer in northern India than in the southern India. See figure 15 for the tree inferred from these 61 individuals. Again, 2 major clusters were observed (clusters A and B), as well as one smaller one (cluster C), and one branch including only one individual. Individuals from all castes were again represented within each cluster.

The Indian population related to other global populations

The Neighbor-Joining tree inferred for 272 sequences from individuals of Indian, European, Middle Eastern and African descent is shown in figure 16. Out of 272 samples, 166 different haplotypes were found. The pygmies form a separate cluster except for two individuals of non-Pygmy origin which are found within this cluster. These two individuals are of Middle-Eastern descent (exact origin unknown) and of European (Sardinian) descent.

The Indian subjects cluster with individuals of both European, Middle Eastern descent. The lower cluster (cluster A) is unusual in that it contains individuals of European and Middle Eastern descent, but only 6 Indians out of 47 individuals. This represents a statistically significant deviation from expectation at the p < .01 level (χ² = 52.46, 1 d.f.). Four out of these 6 Indian individuals are from northern India (2 from Uttar Pradesh, 1 from Punjab, and one from Madhya Pradesh). Two are of southern origin (Kerala and Tamil Nadu). Both southern
Figure 15: Phylogenetic tree - North Indians only, based on caste.

Phylogenetic tree inferred for 61 Indians of northern origin sequenced in this study, plus 14 Pygmies (Vigilant et al., 1991). Symbols correspond to the following subgroups:

- $\Delta$ Pygmies
- $\diamondsuit$ Brahmin
- $\heartsuit$ Kshatriya
- $o$ Vaisya
- $\diamond$ Muslim
- $x$ Christian
Figure 16: Phylogenetic tree - Indians and other human populations.

Phylogenetic inferred for 272 individuals of various descents, including 162 Indians. Symbols correspond to the following subgroups. Each symbol represents one individual:

Δ Pygmy
♦ Indian
• European
○ Middle Eastern
Indians belong to the Brahmin caste. Of the northern individuals, there are one Brahmin, one Kshatriya, one Vaisya and one unknown. The numbers of Indians in clusters C and D are greater than expected at a statistically significant level ($p < .01$, $\chi^2 = 13.79$, 1 d.f., for cluster C; and $p < .01$, $\chi^2 = 22.34$, 1 d.f., for cluster D).

Bootstrap analysis reveals that the tree is in fact a starlike cluster, as only one of the branches is significant at the 95% confidence level. Twenty-one branches are found in > 50 out of 100 bootstraps. Six out of these are within the Pygmy cluster. The branch supporting the two Caucasian individuals found within the Pygmy cluster is found in 50 out of 100 bootstraps. Most of the branches found in > 50% of bootstraps are external branches. None of the internal branching points occur in > 50% of bootstraps, indicating that the phylogeny is indeed starlike.

**Distributions of Pairwise Sequence Differences**

The mean number of pairwise sequence differences within a subgroup can give us an indication of the relative amounts of genetic diversity within the groups. See table 12 for a summary of the mean number of sequence differences within the different religious subgroups and castes. The mean number of differences within the whole population is 3.99 over 195 bp (~2%), indicating that the average difference between two sequences is close to 4 basepairs out of 195 bp. Of the three castes examined in this study, the mean number of pairwise differences is
Table 12: Mean sequence differences within different castes and religions. Sample mean ($\bar{x}$) and sample variance ($s^2$) are given for various castes and religious subgroups of the population. The coefficient of dispersion (C.D. = $s^2 / \bar{x}$) is computed as a rough indication of each distribution’s fit to a Poisson distribution (Sokal and Rohlf, 1981).

<table>
<thead>
<tr>
<th>Caste/Religion</th>
<th>$\bar{x}$</th>
<th>$s^2$</th>
<th>C.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Individuals</td>
<td>3.99</td>
<td>3.65</td>
<td>0.92</td>
</tr>
<tr>
<td>Brahmin</td>
<td>4.43</td>
<td>3.30</td>
<td>0.74</td>
</tr>
<tr>
<td>Kshatriya</td>
<td>3.46</td>
<td>3.60</td>
<td>1.04</td>
</tr>
<tr>
<td>Vaisya</td>
<td>3.97</td>
<td>3.23</td>
<td>0.81</td>
</tr>
<tr>
<td>Christian</td>
<td>2.75</td>
<td>3.53</td>
<td>1.28</td>
</tr>
<tr>
<td>Muslim</td>
<td>5.27</td>
<td>1.78</td>
<td>0.34</td>
</tr>
</tbody>
</table>
highest in Brahmins (4.43), indicating greater genetic diversity within this group. The mean number of pairwise sequence differences was the greatest within the Muslim group (5.27), and the smallest within the Christian group (2.75).

Distributions of pairwise sequence differences are shown in figure 17 for all 84 samples as well as for subsets of samples divided by caste and religion. The distribution of pairwise differences for the whole population as well as for the Brahmins and the Kshatriyas appear to be unimodal, while the distribution for the Vaisya is multimodal. Both the distributions for Christians and for Muslims are multimodal. It is important to bear in mind however, that the sample sizes of these subgroups are very small (7 and 6 individuals respectively).

Under a model of non-constant population growth, distributions of pairwise differences are expected to emulate a Poisson distribution. As the sample mean and sample variance are equal under the Poisson model, a rough estimate of the fit of a distribution to a Poisson curve can be obtained by computing the coefficient of dispersion (C.D.), which should be equal to approximately 1 for a Poisson distribution. Distribution means, variances and coefficients of differentiation are given in table 12. The coefficient of dispersion for the Kshatriya subgroup is equal to 1.04, suggesting that this distribution may follow a Poissonian curve. The remainder of the values for coefficient of dispersion varied more from 1.

In order to determine whether these distributions differed significantly from the Poisson model expected under previous population expansion events, Poisson curves were generated for each subgroup given the sample mean. A
Figure 17: Distributions of frequencies of pairwise sequence differences.

Frequency of the occurrence of each value of pairwise differences is plotted for the total population as well as for each subgroup of caste or religion. Graphs have been normalized by setting the area of each curve equal to 1.
Goodness-of-fit test was employed to compare the observed number of events under each value to those expected under a Poisson distribution. Results of these tests are given in table 13. For the Muslim and Christian subgroups, sample sizes below 30 comparisons make the use of Chi-square analysis unreliable. For the Chi-square Goodness-of-fit tests, all distributions except for that of the Kshatriya group differed quite significantly from expectation under a Poisson model. We therefore cannot reject the null hypothesis that the distribution for the Kshatriya subgroup follows a Poisson model.

Polymorphic Sites

By looking at polymorphic sites separately, one can gain information about a population which may be obscured by the larger degree of minor variations dealt with by analysis of sequences alone.

A total of 8 out of 54 variable sites were polymorphic at the 5% level. The geographical distributions of these substitutions are given in figures 18-25. Chi-square analyses of significance on the basis of northern vs. southern location are given in table 14. The distributions were significant at the p < .05 level for sites 16126, 16223 and 16318 and at the p < .01 level for site 16304. Percentages of individuals carrying substitutions at these sites were computed for the whole Indian population, the northern and southern subpopulations, and for various other human populations. See tables 15-22 for summaries of these results. The frequencies in these tables were computed for individuals of African,
Table 13: Test of Goodness-of-fits to Poisson distributions. $\chi^2$ values are given for tests of Goodness-of-fit of the observed frequencies of each category of number of pairwise differences to those predicted under a Poisson model given the sample means shown in table 12.

<table>
<thead>
<tr>
<th>Caste/Religion</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Population</td>
<td>67.43</td>
<td>9</td>
<td>$p &lt; .00000$</td>
</tr>
<tr>
<td>Brahmin</td>
<td>30.62</td>
<td>7</td>
<td>$p &lt; .00007$</td>
</tr>
<tr>
<td>Kshatriya</td>
<td>3.11</td>
<td>4</td>
<td>$p = .53885$</td>
</tr>
<tr>
<td>Vaisya</td>
<td>35.62</td>
<td>6</td>
<td>$p &lt; .00000$</td>
</tr>
</tbody>
</table>
Table 14: Chi-square analysis of substitution frequencies in southern vs. northern Indian individuals, computed for 8 polymorphic sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>Frequency of substitution</th>
<th>Total Population</th>
<th>$\chi^2$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Northern Indians</td>
<td>Southern Indians</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16126</td>
<td>11/71</td>
<td>7/111</td>
<td>118/182</td>
<td>4.10</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>6%</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>16129</td>
<td>11/71</td>
<td>14/111</td>
<td>125/182</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>13%</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>16172</td>
<td>5/71</td>
<td>15/111</td>
<td>20/182</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td>7%</td>
<td>14%</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td>16223</td>
<td>38/71</td>
<td>76/111</td>
<td>114/182</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>54%</td>
<td>68%</td>
<td>63%</td>
<td></td>
</tr>
<tr>
<td>16304</td>
<td>9/71</td>
<td>2/110</td>
<td>11/181</td>
<td>8.91</td>
</tr>
<tr>
<td></td>
<td>9%</td>
<td>2%</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>16309</td>
<td>5/71</td>
<td>3/110</td>
<td>8/181</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>5%</td>
<td>3%</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>16311</td>
<td>9/71</td>
<td>11/109</td>
<td>120/180</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>13%</td>
<td>10%</td>
<td>11%</td>
<td></td>
</tr>
<tr>
<td>16318</td>
<td>7/71</td>
<td>2/109</td>
<td>9/179</td>
<td>5.95</td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>2%</td>
<td>5%</td>
<td></td>
</tr>
</tbody>
</table>
Asian and Caucasian descent, and independently for individuals of European, Middle Eastern, East Asian and Native American descent. For the latter compilations, only those sequences for whom geographical origin was known beyond major race were included.

The C→T transition at site 16223 is found in 63% of Indian individuals. It is found in 68% of south Indians, but only in 54% of north Indians. This same substitution is rare in Europeans (9%), but quite prevalent in both Asians and Africans (62% and 78% respectively). The T substitution occurs within members of all castes and religions. However, it is found in fewer members of the Scheduled Castes than expected, and in more members of the Vaisya caste than expected ($\chi^2 = 4.24$ and 4.88 respectively, d.f. = 1, $p < .05$). Individuals with a C at this site are widely distributed over the northern part of India, with no specific clustering of this substitution within any one region.

The T→C transition at site 16126 occurs in 10% of Indian individuals. It is seen in 15% of north Indians and in 6% of south Indians. All 7 south Indians with the C substitution belong to the Brahmin caste. Of the 11 north Indians with the C substitution, 6 belong to the Brahmin caste, 3 to the Kshatriya caste and 1 is Muslim. The caste of the remaining individual is unknown. Thus, a total of 13 out of the 18 individuals with this substitution belong to the Brahmin caste. The northern individuals with this substitution are found mainly in Uttar Pradesh and West Bengal, as well as one individual from Maharashtra. The frequency of this substitution in Caucasians (of European and Middle Eastern descent) is 8%. The
frequency in Asians is quite low (2%) and in Africans it is moderate at 6%.

At site 16304, the T→C transition is found in 6% of Indians. 13% of north Indians have the C substitution, while only 2% of south Indians have the same substitution. Both south Indians with the substitution belong to the Brahmin caste. Of the north Indians, 1 is a Brahmin, 1 a Kshatriya, 3 are Vaisyas, 1 is Christian 2 are Muslim and 1 is unknown. As can be seen in figure 22, the geographic locations of these individuals are also quite widespread. The occurrence of this substitution in both Vaisyas and Muslims is greater than expected. The T→C substitution occurs in 5% of Caucasians, 14% of Asians and 0% of Africans.

Site 16318 is unique among the more polymorphic sites, in that the substitution in this case is a transversion. The A→T transversion occurs in 10% of Indians (9 individuals). One additional individual has an A→C transversion. The prevalence in north Indians is 10%, and in south Indians 2%. Both south Indians belong to the Brahmin caste. Of the north Indians, 3 are Brahmins, 1 is a Kshatriya, 1 a Vaisya and 2 are unknown. The geographic origins of these individuals are all in the north of India, with individuals being from Uttar Pradesh, Punjab, Bihar, West-Bengal and Nepal. The A→C transversion is found in a Muslim individual from Kashmir.

This site is of interest in that this particular transversion appears to be quite rare within other populations. It was not found in any of the 212 Caucasian or African samples. It was seen in only 2% of Asians, 2 individuals of Japanese origin. In both individuals, the substitution was an A→T transversion.
Figures 18 - 25: Each circle represents one individual. Squares represent individuals for whom geographic location is not known more precisely than the province. Triangles represent the origin of 98 previously published south Indian samples (Mountain et al., 1995). The 7 samples represented by the triangle in Kerala are all samples of tribal origin. For some of the published sequences from Karnataka, the particular site in question was unresolved, and these sequences were not included in the analysis.

Figure 18: Geographic distribution of substitutions at site 16126a.

* base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16126
- T
- C
Table 15: Frequencies of T→C substitution at polymorphic site 16126 for different human populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>T</th>
<th>C</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian(^a)</td>
<td>60</td>
<td>11</td>
<td>71</td>
<td>15%</td>
</tr>
<tr>
<td>South Indian(^a)</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td>Karnataka(^b)</td>
<td>84</td>
<td>7</td>
<td>91</td>
<td>8%</td>
</tr>
<tr>
<td>Kerala(^b)</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian(^c)</td>
<td>60</td>
<td>11</td>
<td>71</td>
<td>15%</td>
</tr>
<tr>
<td>Combined South Indian(^c)</td>
<td>104</td>
<td>7</td>
<td>111</td>
<td>6%</td>
</tr>
<tr>
<td>Indian</td>
<td>164</td>
<td>18</td>
<td>182</td>
<td>10%</td>
</tr>
<tr>
<td>European</td>
<td>88</td>
<td>6</td>
<td>94</td>
<td>6%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>39</td>
<td>3</td>
<td>42</td>
<td>7%</td>
</tr>
<tr>
<td>East Asian</td>
<td>92</td>
<td>2</td>
<td>94</td>
<td>2%</td>
</tr>
<tr>
<td>Native American</td>
<td>72</td>
<td>0</td>
<td>72</td>
<td>0%</td>
</tr>
<tr>
<td>Caucasian(^d)</td>
<td>148</td>
<td>13</td>
<td>161</td>
<td>8%</td>
</tr>
<tr>
<td>Asian(^b)</td>
<td>92</td>
<td>2</td>
<td>94</td>
<td>2%</td>
</tr>
<tr>
<td>Native American</td>
<td>72</td>
<td>0</td>
<td>72</td>
<td>0%</td>
</tr>
<tr>
<td>African</td>
<td>45</td>
<td>3</td>
<td>48</td>
<td>6%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>357</td>
<td>18</td>
<td>375</td>
<td>5%</td>
</tr>
</tbody>
</table>

\(^a\) From present study.
\(^b\) From Mountain et al., 1995.
\(^c\) Combined data from present study and Mountain et al., 1995.
\(^d\) Includes individuals of European, Middle Eastern and North American descent.
\(^*\) Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Figure 19: Geographic distribution of substitutions at site 16129a.

a base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16129

- ○ G
- ● A
Table 16: Frequencies of G→A substitution at polymorphic site 16129 for different human populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>G</th>
<th>A</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian\textsuperscript{a}</td>
<td>60</td>
<td>11</td>
<td>71</td>
<td>15%</td>
</tr>
<tr>
<td>South Indian\textsuperscript{a}</td>
<td>11</td>
<td>2</td>
<td>13</td>
<td>15%</td>
</tr>
<tr>
<td>Karnataka\textsuperscript{b}</td>
<td>79</td>
<td>12</td>
<td>91</td>
<td>13%</td>
</tr>
<tr>
<td>Kerala\textsuperscript{b}</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian\textsuperscript{c}</td>
<td>60</td>
<td>11</td>
<td>71</td>
<td>15%</td>
</tr>
<tr>
<td>Combined South Indian\textsuperscript{c}</td>
<td>97</td>
<td>14</td>
<td>111</td>
<td>13%</td>
</tr>
<tr>
<td>Indian</td>
<td>157</td>
<td>25</td>
<td>182</td>
<td>14%</td>
</tr>
<tr>
<td>European</td>
<td>94</td>
<td>0</td>
<td>94</td>
<td>0%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>42</td>
<td>0</td>
<td>42</td>
<td>0%</td>
</tr>
<tr>
<td>East Asian</td>
<td>76</td>
<td>18</td>
<td>94</td>
<td>19%</td>
</tr>
<tr>
<td>Native American</td>
<td>71</td>
<td>1</td>
<td>72</td>
<td>1%</td>
</tr>
<tr>
<td>Caucasian\textsuperscript{d}</td>
<td>162</td>
<td>1</td>
<td>163</td>
<td>1%</td>
</tr>
<tr>
<td>Asian\textsuperscript{e}</td>
<td>76</td>
<td>18</td>
<td>94</td>
<td>19%</td>
</tr>
<tr>
<td>Native American</td>
<td>71</td>
<td>1</td>
<td>72</td>
<td>1%</td>
</tr>
<tr>
<td>African</td>
<td>32</td>
<td>17</td>
<td>49</td>
<td>35%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>341</td>
<td>37</td>
<td>378</td>
<td>10%</td>
</tr>
</tbody>
</table>

\textsuperscript{a} From present study.
\textsuperscript{b} From Mountain \textit{et al.}, 1995.
\textsuperscript{c} Combined data from present study and Mountain \textit{et al.}, 1995.
\textsuperscript{d} Includes individuals of European, Middle Eastern and North American descent.
\textsuperscript{e} Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Figure 20: Geographic distribution of substitutions at site 16172a.

• base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16172

- O T
- • C
Table 17: Frequencies of T→C substitution at polymorphic site 16172 for different human populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>T</th>
<th>C</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian(^a)</td>
<td>66</td>
<td>5</td>
<td>71</td>
<td>7%</td>
</tr>
<tr>
<td>South Indian(^a)</td>
<td>12</td>
<td>1</td>
<td>13</td>
<td>8%</td>
</tr>
<tr>
<td>Karnataka(^b)</td>
<td>77</td>
<td>14</td>
<td>91</td>
<td>15%</td>
</tr>
<tr>
<td>Kerala(^b)</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian(^c)</td>
<td>66</td>
<td>5</td>
<td>71</td>
<td>7%</td>
</tr>
<tr>
<td>Combined South Indian(^c)</td>
<td>96</td>
<td>15</td>
<td>111</td>
<td>14%</td>
</tr>
<tr>
<td>Indian</td>
<td>162</td>
<td>20</td>
<td>182</td>
<td>11%</td>
</tr>
<tr>
<td>European</td>
<td>94</td>
<td>0</td>
<td>94</td>
<td>0%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>42</td>
<td>0</td>
<td>42</td>
<td>0%</td>
</tr>
<tr>
<td>East Asian</td>
<td>85</td>
<td>9</td>
<td>94</td>
<td>10%</td>
</tr>
<tr>
<td>Native American</td>
<td>68</td>
<td>4</td>
<td>72</td>
<td>6%</td>
</tr>
<tr>
<td>Caucasian(^d)</td>
<td>160</td>
<td>3</td>
<td>163</td>
<td>2%</td>
</tr>
<tr>
<td>Asian(^e)</td>
<td>85</td>
<td>9</td>
<td>94</td>
<td>10%</td>
</tr>
<tr>
<td>Native American</td>
<td>68</td>
<td>4</td>
<td>72</td>
<td>6%</td>
</tr>
<tr>
<td>African</td>
<td>43</td>
<td>6</td>
<td>49</td>
<td>12%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>356</td>
<td>22</td>
<td>378</td>
<td>6%</td>
</tr>
</tbody>
</table>

\(^a\) From present study.
\(^b\) From Mountain et al., 1995.
\(^c\) Combined data from present study and Mountain et al., 1995.
\(^d\) Includes individuals of European, Middle Eastern and North American descent.
\(^e\) Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
**Figure 21:** Geographic distribution of substitutions at site 16223\textsuperscript{a}.

- base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16223

- C
- T
Table 18: Frequencies of C→T substitution at polymorphic site 16223 for different human populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>C</th>
<th>T</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian&lt;sup&gt;a&lt;/sup&gt;</td>
<td>33</td>
<td>38</td>
<td>71</td>
<td>54%</td>
</tr>
<tr>
<td>South Indian&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2</td>
<td>11</td>
<td>13</td>
<td>85%</td>
</tr>
<tr>
<td>Karnataka&lt;sup&gt;b&lt;/sup&gt;</td>
<td>28</td>
<td>63</td>
<td>91</td>
<td>69%</td>
</tr>
<tr>
<td>Kerala&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>29%</td>
</tr>
<tr>
<td>Combined North Indian&lt;sup&gt;c&lt;/sup&gt;</td>
<td>33</td>
<td>38</td>
<td>71</td>
<td>54%</td>
</tr>
<tr>
<td>Combined South Indian&lt;sup&gt;c&lt;/sup&gt;</td>
<td>35</td>
<td>76</td>
<td>111</td>
<td>68%</td>
</tr>
<tr>
<td>Indian</td>
<td>68</td>
<td>114</td>
<td>182</td>
<td>63%</td>
</tr>
<tr>
<td>European</td>
<td>86</td>
<td>8</td>
<td>94</td>
<td>9%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>31</td>
<td>11</td>
<td>42</td>
<td>26%</td>
</tr>
<tr>
<td>East Asian</td>
<td>36</td>
<td>58</td>
<td>94</td>
<td>62%</td>
</tr>
<tr>
<td>Native American</td>
<td>14</td>
<td>58</td>
<td>72</td>
<td>81%</td>
</tr>
<tr>
<td>Caucasian&lt;sup&gt;d&lt;/sup&gt;</td>
<td>140</td>
<td>22</td>
<td>162</td>
<td>14%</td>
</tr>
<tr>
<td>Asian&lt;sup&gt;e&lt;/sup&gt;</td>
<td>36</td>
<td>58</td>
<td>94</td>
<td>62%</td>
</tr>
<tr>
<td>Native American</td>
<td>14</td>
<td>58</td>
<td>72</td>
<td>81%</td>
</tr>
<tr>
<td>African</td>
<td>11</td>
<td>38</td>
<td>49</td>
<td>78%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>201</td>
<td>176</td>
<td>377</td>
<td>47%</td>
</tr>
</tbody>
</table>

<sup>a</sup> From present study.
<sup>b</sup> From Mountain <i>et al.</i>, 1995.
<sup>c</sup> Combined data from present study and Mountain <i>et al.</i>, 1995.
<sup>d</sup> Includes individuals of European, Middle Eastern and North American descent.
<sup>e</sup> Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Figure 22: Geographic distribution of substitutions at site 16304a.

- base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16304

○ T
● C
Table 19: Frequencies of T→C substitution at polymorphic site 16304 for different human populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>T</th>
<th>C</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian\textsuperscript{a}</td>
<td>62</td>
<td>9</td>
<td>71</td>
<td>13%</td>
</tr>
<tr>
<td>South Indian\textsuperscript{a}</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>0%</td>
</tr>
<tr>
<td>Karnataka\textsuperscript{b}</td>
<td>88</td>
<td>2</td>
<td>90</td>
<td>2%</td>
</tr>
<tr>
<td>Kerala\textsuperscript{b}</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian\textsuperscript{c}</td>
<td>62</td>
<td>9</td>
<td>71</td>
<td>13%</td>
</tr>
<tr>
<td>Combined South Indian\textsuperscript{c}</td>
<td>108</td>
<td>2</td>
<td>110</td>
<td>2%</td>
</tr>
<tr>
<td>Indian</td>
<td>170</td>
<td>11</td>
<td>181</td>
<td>6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>T</th>
<th>C</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>European</td>
<td>91</td>
<td>3</td>
<td>94</td>
<td>3%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>42</td>
<td>0</td>
<td>42</td>
<td>0%</td>
</tr>
<tr>
<td>East Asian</td>
<td>81</td>
<td>13</td>
<td>94</td>
<td>14%</td>
</tr>
<tr>
<td>Native American</td>
<td>70</td>
<td>2</td>
<td>72</td>
<td>3%</td>
</tr>
<tr>
<td>Caucasian\textsuperscript{d}</td>
<td>155</td>
<td>8</td>
<td>163</td>
<td>5%</td>
</tr>
<tr>
<td>Asian\textsuperscript{*}</td>
<td>81</td>
<td>13</td>
<td>94</td>
<td>14%</td>
</tr>
<tr>
<td>Native American</td>
<td>70</td>
<td>2</td>
<td>72</td>
<td>3%</td>
</tr>
<tr>
<td>African</td>
<td>49</td>
<td>0</td>
<td>49</td>
<td>0%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>355</td>
<td>23</td>
<td>378</td>
<td>6%</td>
</tr>
</tbody>
</table>

\textsuperscript{a} From present study.
\textsuperscript{b} From Mountain \textit{et al.}, 1995.
\textsuperscript{c} Combined data from present study and Mountain \textit{et al.}, 1995.
\textsuperscript{d} Includes individuals of European, Middle Eastern and North American descent.
\textsuperscript{*} Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Figure 23: Geographic distribution of substitutions at site 16309.

- base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16309

- A
- G
<table>
<thead>
<tr>
<th>Population</th>
<th>G</th>
<th>A</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian&lt;sup&gt;a&lt;/sup&gt;</td>
<td>66</td>
<td>5</td>
<td>71</td>
<td>7%</td>
</tr>
<tr>
<td>South Indian&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12</td>
<td>1</td>
<td>13</td>
<td>8%</td>
</tr>
<tr>
<td>Karnataka&lt;sup&gt;b&lt;/sup&gt;</td>
<td>88</td>
<td>2</td>
<td>90</td>
<td>2%</td>
</tr>
<tr>
<td>Kerala&lt;sup&gt;b&lt;/sup&gt;</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian&lt;sup&gt;c&lt;/sup&gt;</td>
<td>66</td>
<td>5</td>
<td>71</td>
<td>7%</td>
</tr>
<tr>
<td>Combined South Indian&lt;sup&gt;c&lt;/sup&gt;</td>
<td>107</td>
<td>3</td>
<td>110</td>
<td>3%</td>
</tr>
<tr>
<td>Indian</td>
<td>173</td>
<td>8</td>
<td>181</td>
<td>4%</td>
</tr>
<tr>
<td>European</td>
<td>93</td>
<td>1</td>
<td>94</td>
<td>1%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>38</td>
<td>4</td>
<td>42</td>
<td>10%</td>
</tr>
<tr>
<td>East Asian</td>
<td>92</td>
<td>2</td>
<td>94</td>
<td>2%</td>
</tr>
<tr>
<td>Native American</td>
<td>72</td>
<td>0</td>
<td>72</td>
<td>0%</td>
</tr>
<tr>
<td>Caucasian&lt;sup&gt;d&lt;/sup&gt;</td>
<td>158</td>
<td>5</td>
<td>163</td>
<td>3%</td>
</tr>
<tr>
<td>Asian&lt;sup&gt;e&lt;/sup&gt;</td>
<td>92</td>
<td>2</td>
<td>94</td>
<td>2%</td>
</tr>
<tr>
<td>Native American</td>
<td>72</td>
<td>0</td>
<td>72</td>
<td>0%</td>
</tr>
<tr>
<td>African</td>
<td>48</td>
<td>1</td>
<td>49</td>
<td>2%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>370</td>
<td>8</td>
<td>378</td>
<td>2%</td>
</tr>
</tbody>
</table>

<sup>a</sup> From present study.
<sup>b</sup> From Mountain <i>et al.</i>, 1995.
<sup>c</sup> Combined data from present study and Mountain <i>et al.</i>, 1995.
<sup>d</sup> Includes individuals of European, Middle Eastern and North American descent.
<sup>e</sup> Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Figure 24: Geographic distribution of substitutions at site 16311a.

- base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16311

- T
- C
<table>
<thead>
<tr>
<th>Population</th>
<th>T</th>
<th>C</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian\textsuperscript{a}</td>
<td>62</td>
<td>9</td>
<td>71</td>
<td>13%</td>
</tr>
<tr>
<td>South Indian\textsuperscript{a}</td>
<td>12</td>
<td>1</td>
<td>13</td>
<td>8%</td>
</tr>
<tr>
<td>Karnataka\textsuperscript{b}</td>
<td>79</td>
<td>10</td>
<td>89</td>
<td>11%</td>
</tr>
<tr>
<td>Kerala\textsuperscript{b}</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian\textsuperscript{c}</td>
<td>62</td>
<td>9</td>
<td>71</td>
<td>13%</td>
</tr>
<tr>
<td>Combined South Indian\textsuperscript{c}</td>
<td>98</td>
<td>11</td>
<td>109</td>
<td>10%</td>
</tr>
<tr>
<td>Indian</td>
<td>160</td>
<td>20</td>
<td>180</td>
<td>11%</td>
</tr>
<tr>
<td>European</td>
<td>89</td>
<td>5</td>
<td>94</td>
<td>5%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>41</td>
<td>1</td>
<td>42</td>
<td>2%</td>
</tr>
<tr>
<td>East Asian</td>
<td>82</td>
<td>12</td>
<td>94</td>
<td>13%</td>
</tr>
<tr>
<td>Native American</td>
<td>67</td>
<td>5</td>
<td>72</td>
<td>7%</td>
</tr>
<tr>
<td>Caucasian\textsuperscript{d}</td>
<td>151</td>
<td>12</td>
<td>163</td>
<td>7%</td>
</tr>
<tr>
<td>Asian\textsuperscript{e}</td>
<td>82</td>
<td>12</td>
<td>94</td>
<td>13%</td>
</tr>
<tr>
<td>Native American</td>
<td>67</td>
<td>5</td>
<td>72</td>
<td>7%</td>
</tr>
<tr>
<td>African</td>
<td>23</td>
<td>26</td>
<td>49</td>
<td>53%</td>
</tr>
<tr>
<td>Total (other populations)</td>
<td>323</td>
<td>55</td>
<td>378</td>
<td>15%</td>
</tr>
</tbody>
</table>

\textsuperscript{a} From present study.
\textsuperscript{b} From Mountain \textit{et al.}, 1995.
\textsuperscript{c} Combined data from present study and Mountain \textit{et al.}, 1995.
\textsuperscript{d} Includes individuals of European, Middle Eastern and North American descent.
\textsuperscript{e} Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Figure 25: Geographic distribution of substitutions at site 16318a.

* base number according to reference sequence (Anderson et al., 1981)
Polymorphic Sites

Site 16318

- A
- T
- C
Table 22: Frequencies of A→T substitution at polymorphic site 16318 for different human populations.

<table>
<thead>
<tr>
<th>Population</th>
<th>A</th>
<th>T</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Indian(^a)</td>
<td>63</td>
<td>7</td>
<td>70</td>
<td>10%</td>
</tr>
<tr>
<td>South Indian(^a)</td>
<td>12</td>
<td>1</td>
<td>13</td>
<td>8%</td>
</tr>
<tr>
<td>Karnataka(^b)</td>
<td>88</td>
<td>1</td>
<td>89</td>
<td>1%</td>
</tr>
<tr>
<td>Kerala(^b)</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>0%</td>
</tr>
<tr>
<td>Combined North Indian(^c)</td>
<td>63</td>
<td>7</td>
<td>70</td>
<td>10%</td>
</tr>
<tr>
<td>Combined South Indian(^c)</td>
<td>107</td>
<td>2</td>
<td>109</td>
<td>2%</td>
</tr>
<tr>
<td>Indian</td>
<td>170</td>
<td>9</td>
<td>179</td>
<td>5%</td>
</tr>
<tr>
<td>European</td>
<td>94</td>
<td>0</td>
<td>94</td>
<td>0%</td>
</tr>
<tr>
<td>Middle Eastern</td>
<td>42</td>
<td>0</td>
<td>42</td>
<td>0%</td>
</tr>
<tr>
<td>East Asian</td>
<td>92</td>
<td>2</td>
<td>94</td>
<td>2%</td>
</tr>
<tr>
<td>Native American</td>
<td>72</td>
<td>0</td>
<td>72</td>
<td>0%</td>
</tr>
<tr>
<td>Caucasian(^d)</td>
<td>163</td>
<td>0</td>
<td>163</td>
<td>0%</td>
</tr>
<tr>
<td>Asian(^b)</td>
<td>92</td>
<td>2</td>
<td>94</td>
<td>2%</td>
</tr>
<tr>
<td>Native American</td>
<td>72</td>
<td>0</td>
<td>72</td>
<td>0%</td>
</tr>
<tr>
<td>African</td>
<td>49</td>
<td>0</td>
<td>49</td>
<td>0%</td>
</tr>
<tr>
<td><strong>Total (other populations)</strong></td>
<td>376</td>
<td>2</td>
<td>378</td>
<td>1%</td>
</tr>
</tbody>
</table>

\(^a\) From present study.
\(^b\) From Mountain *et al.*, 1995.
\(^c\) Combined data from present study and Mountain *et al.*, 1995.
\(^d\) Includes individuals of European, Middle Eastern and North American descent.
\(^e\) Includes individuals of Japanese, Indonesian, Korean, Chinese, Filipino and Papua New Guinean descent.
Discussion

Phylogenetic Inferences

Phylogenetic trees inferred from sequences of a 195 bp region of non-coding mitochondrial DNA do not demonstrate any distinct clustering with respect to either geographical origin, language or caste.

Geography

There was no significant degree of clustering with respect to geographical origin, as based on province of origin. This may be the result of a number of factors. The sampling distribution includes mainly individuals of urban descent. As such, these individuals or their recent ancestors may have moved from their initial region of origin. It must be noted however, that our assignment of geographic origins were based on the place of birth of the donor's mother, in order to minimize the effects of recent migration. The lack of clustering may also be the result of a significant amount of gene flow between regions, resulting in the loss of any region specific haplotypes. Alternatively, there may not have been sufficient time since the settlement of the ancestors of the various regions for the establishment of any significant sequence differentiation. In order to better address the questions of geographic differentiation, especially with respect to the historic settlement of the different regions, it would be of interest to obtain samples
of tribal origins.

Other human studies which have revealed distinct clustering on the basis of geography, such as those that separate members of the Amerindian population into four distinct clusters, each thought to represent a separate ancestral mtDNA lineage (Horai et al., 1993), may do so as a result the following: a) The founding populations are thought to have undergone a bottleneck prior to re-expansion, at which time the number of mtDNA lineages in each population was dramatically reduced, or b) These populations have been separated for a longer period of time, on the order of 10000-20000 years. In comparison, the majority of the major influential invasions into India occurred within the last 3500 or 4000 years. Providing that the various invaders were of completely disparate genetic heritages, one may be able to trace their influences on Indian populations using genetic techniques. However, if they were all of similar enough descent, mtDNA sequence analysis would be of little use in reconstructing foreign influences.

Language

Using language as an indicator of an individual's heritage should be done with care, as there are exceptions to the trend of linkage of language and heritage. Cases of disparity may be due to either the replacement of language, where people give up their language in favour of a new one, often as a result of foreign rule, or to the replacement of genes as the result of potentially gradual population admixture. A population may carry a genetic signature indicating one heritage, but due to language replacement as a result of various possible factors, speak a
language belonging to a completely different family.

Exclusive statements should also not be made with regard to the role that language plays in affecting any of the genetic structures seen within the Indian population, as language in India is very closely linked to geographical location, and presumably to historical events and migrations as well. Therefore the significance of Dravidian language distribution may be better described as being based on Dravidian peoples.

In the phylogenetic trees inferred for spoken language, there was a significantly smaller than expected representation of Dravidian language speakers in one of the main clusters. If there is a genetic separation on the basis of language, this cluster may represent an Indo-Iranian cluster, containing mainly individuals of Indo-Aryan descent. The occurrence of speakers of both Indo-Iranian and of Dravidian language families in the remainder of the clusters may represent admixture of Dravidian and Indo-European based peoples. Alternatively, these results would also be consistent with the theory that Dravidian language family speakers were once spread over the whole of India, and were absorbed into the later Indo-Aryan culture. In this case some Indo-European language family speakers may in fact be of Dravidian heritage, as a result of language replacement.

Previous studies using gene frequency data have shown the existence of genetic clines at dialect boundaries which did not correspond to geographical boundaries as well (Barbujani, 1991). It is possible that his technique is more
sensitive for detecting the relationship of language and genetics. The use of RFLP frequencies may prove a superior technique for addressing this question.

**Caste**

With regard to caste, the phylogenetic trees inferred from these samples also exhibited no significant degree of clustering with respect to caste. This lack of clustering related to caste has been observed in mtDNA sequence based studies done on populations from Karnataka and from Andhra Pradesh (Bamshad *et al.*, 1994; Mountain *et al.*, 1995). Mountain's (1995) inclusion of individuals of both the highest and the lowest caste failed to reveal any significant clustering, indicating at least some degree of gene flow between even highly separated castes. Branches separating members of the two castes were however more robust, as indicated by bootstrap analysis, suggesting a recent lack of gene flow. A similar result was observed in gene frequency studies on Indian populations in Maharashtra (Malhotra *et al.*, 1978).

Of interest in the phylogenetic tree inferred with respect to caste, was the under-representation of members of the Vaisya caste within one of the clusters. Within the hierarchy of the caste system, the Vaisya caste is below both Brahmins and Kshatriyas. It is possible that this cluster represents mainly members of the higher castes. Both the Christians and Muslims found in this cluster may be converts from higher Hindu castes as well. A study of northern Indian populations has also found the Vaisya caste to be a unique group, separated from the remainder of the castes by the maximum genetic distance (Kamboh, 1984).
With respect to lack of resolution of caste differences within the phylogenetic trees, there are a number of possible interpretations. One of these would be that the strict restrictions on intermarriage between castes are not practiced as stringently as expected, leading to some degree of gene flow between castes. Marjoram and Donnelly (1994) showed that amounts of migration on the order of 0.01 individual/generation will lead to a unimodal Poisson distribution of pairwise differences, which often corresponds to starlike phylogenies without resolvable branching orders at the centres.

Another possibility relates to the age of the caste system. When dealing with mtDNA lineages, it is important to bear in mind that although the lineages of mtDNA can be separated as far back as the origin of the population, the separation of lineages may well be more ancient (Avise et al., 1984). In other words, the separation of lineages may be older than the caste system, resulting in representation of all or most lineages within each caste. The 2000 years during which the caste system has been in existence may not have been sufficient to allow for the dying out or fixation of enough mtDNA lineages within the different castes to create any discernable variation.

In order to allay any such effects resulting from inclusion of the southern samples (where the caste system has not been practiced for nearly as long), a phylogeny including only the northern samples was generated. This however showed no more clustering than the phylogeny including both northern and southern sequences, indicating that the southern samples are not skewing the
Some previous genetic studies employing frequencies of blood group markers and gene frequency data did however reveal some degree of genetic separation on the basis of castes (Saha et al., 1992; Walter et al., 1993). This may indicate that sequencing of the mitochondrial D-loop is not a sensitive enough technique to resolve these differences.

When addressing caste, it is important to recall the maternal inheritance of mitochondrial DNA. Analysis based on mtDNA only would be sensitive to any differences between the sexes with respect to movement between castes. Within the Indian caste system, marriage of women into higher status castes is permissible (Cavalli-Sforza et al., 1994). The result of this could be a reasonable amount of flow of genetic material from lower to higher castes. MtDNA studies may exaggerate the discrepancies between castes. A similar study using Y-chromosome markers of a similar degree of variability would be of use in supplying the paternal side of genetic histories, and in revealing if there is indeed an effect of the unisexual transmission of mtDNA, which may be skewing the results.

Robustness of the Use of Phylogenies to Interpret Population Structuring and Genetic Divergences

Templeton's Delta Q test reveals that the phylogenetic trees inferred using the Neighbor-Joining method, and those inferred using the Maximum Parsimony method do not differ statistically from each other with respect to the number of
steps required to generate each one, in all but one case. The use of Neighbor-Joining trees in drawing inferences therefore seems reasonable.

Bootstrap analysis reveals that few of the branches are robust at the 95% confidence level, suggesting that there is little certainty with respect to branching order. This type of occurrence is often the result of a starlike phylogeny, where most branches originate at one point in the tree. The order of branching is difficult to infer in such cases. A starlike phylogeny results from coalescent events occurring within a narrow time frame. This may be due to a recent exponential growth event, a population bottleneck or founder effect, or natural selection (Marjoram and Donnelly, 1994; Slatkin and Hudson, 1991).

**Population Demography**

The distributions of frequencies of pairwise sequence differences can be used to draw inferences regarding population histories. Under a model of roughly constant population size, distributions tend to be bimodal or multimodal. Populations which have exhibited exponential growth, either suddenly, or over time, will have distributions that are unimodal, and approximately follow a Poisson distribution (Slatkin and Hudson, 1991).

Distributions of pairwise differences of sequences appeared to be unimodal for the entire Indian population sampled in this study, as well as for Brahmins and Kshatriyas. The Vaisya distribution exhibits a multimodal distribution. Only the Kshatriya distribution however, is not significantly different from a Poisson
distribution. The sample sizes for both the Muslim and Christian groups, were too small to allow any conclusive interpretation of distributions. In the study by Mountain (1995), Brahmins exhibited a unimodal distribution, while the Scheduled Caste group exhibited a multimodal distribution. Higher castes therefore may to have undergone rapid expansions in population size, while lower castes such as the Vaisyas and Scheduled Castes have maintained a constant population size. This may be consistent with extra opportunities for expansion granted to members of higher castes due to higher economic stability.

Inferences concerning population growth histories, which are drawn from data based on pairwise comparisons of mtDNA sequences should however be made with caution, as a number of factors can affect mismatch distributions. Unimodal distributions may be the result of any factors that force coalescent events into a narrow period of time. These include natural selection, population bottlenecks or founder effects (Marjoram and Donnelly, 1994; Slatkin and Hudson, 1991). A multimodal distribution may not be the result of constant population size, but rather of an exponentially growing, but subdivided population with little migration occurring between subpopulations. A large initial population size will also result in a multimodal distribution, even if the population undergoes exponential growth at some point (Marjoram and Donnelly, 1994).

An interesting result with respect to religion was observed in the within group sequence variabilities. The large mean number of pairwise sequence differences seen in the Muslim group, in comparison to the very small mean
observed in the Christian group may be an indication that converts to Islam came from a much broader base with respect to caste or geographical origin than did converts to Christianity. This may also be a result of Christians in India tending to be of more recent conversion than Muslims.

**Informative Polymorphic Sites**

The analysis of frequency distributions of substitutions at individual polymorphic sites is another approach to examining population histories. By comparing frequencies of substitutions among different populations, one may be able to draw hypotheses with respect to interrelationships and population histories based on previous invasions and migrations.

Site 16223 was the most informative of the polymorphic sites studied, mainly due to the large degree of polymorphism present at this site within the Indian population. The frequency of the $C\rightarrow T$ substitution in north Indians was significantly higher than in south Indians. At the same time, this substitution is rare in Europeans, but prevalent in both Asians and Africans. As more northern individuals are assumed to be of descent resulting from the Aryan invasions, it is plausible that the lower incidence of the $C\rightarrow T$ substitution in the north, reflects a greater degree of relationship between north Indians and Europeans.

The higher incidence in the south corresponds to a higher incidence in Asian and African individuals. While little is known of the genetic heritage of Dravidian-based, south Indians, these results may indicate some form of
relationship with Africans or Asians. Melton (1995), obtained similar results with respect to the incidence of a 9 bp deletion region previously considered to be Asian-specific (Horai et al., 1993). The frequency of the 9 bp deletion in south Indians was found to be higher (8%) than in north Indians (0%). As the individuals in these studies are not of tribal origin, drawing conclusions based on these data with regard to the original inhabitants of India is not feasible. Both of these results may instead indicate a recent influx of individuals of Asian descent into southern India.

The three remaining sites which showed a statistically significant difference between frequency of substitutions between north and south Indians are not very highly polymorphic sites within the Indian population, occurring in only 10%, 6% and 5% of Indian individuals. As a result of the smaller degree of polymorphism exhibited at these sites, it is difficult to observe any significant trends with respect to north-south differences, or to make comparisons to other populations. It would be more useful to locate more highly polymorphic sites such as site 16223, which may yield more stringent information due to their higher frequency of occurrence. A survey of RFLP sites to locate more such sites would be of great interest. Frequencies of these specific substitutions within different world populations could then be compared to reveal any significant trends which may yield information regarding the interrelationships of these populations. It would also be beneficial to survey more human populations, especially from countries immediately surrounding India, in order to attempt to localize influxes of genetic material into
the country in an effort to attempt to recreate historical migrations.

**Indians Within the Global Human Population**

Indians cluster with Europeans, and individuals of Middle Eastern descent, but tend not to cluster with Africans. This same sort of pattern was found by Mountain (1995). The clustering of Indians with Europeans is not surprising as at least north Indians are believed to be of Indo-Aryan descent. Indians believed to be descended from the previous inhabitants, the Dravidians, are also believed to have undergone a fair amount of admixture with the Indo-Aryan invaders, both during the original waves of migration, and presumably subsequently as well.

The distribution of Indians within clusters was not completely random however, with one cluster containing significantly fewer than expected Indians. The obtainment of more sequences from the areas surrounding the Indian subcontinent may help clarify the basis of this trend, as might a more sensitive genetic technique. When attempting to localize a population within a global phylogeny, it must be remembered that at present, not all human populations have been surveyed for mtDNA D-loop variation, thus creating gaps in potential interrelationships.

**New directions for Future Research**

Phylogenies inferred from different human sequence data appear to give similar results, a lack of resolution at the centre of a star phylogeny. MtDNA may
be not be powerful enough to elucidate the interrelationships between closely related individuals. For the purpose of detecting population substructure of the level expected under the supposedly strict endogamy practiced under the caste system, the technique again may not be sensitive enough, which is important to recall before drawing any sorts of conclusions regarding the effectiveness of the caste system of restricted intermarriage.

Bootstrap analysis reveals that few branches are robust even at the 50% significance level. Therefore, all phylogenies are truly star phylogenies, with the branching order being difficult to resolve at the centre.

It would be beneficial to do more frequency comparisons for different populations with respect to polymorphic sites. A good method for surveying large numbers of these sites would be to determine any RFLP sites corresponding to Indian specific substitutions.

The sampling distribution needs to be improved in order to make the study more complete. Additional samples should be obtained for some of the regions for which there are presently few or no samples. This would allow the testing of caste differences within the same geographic area, in order to remove any noise possibly due to geographical genetic differences. As well, having a greater sample size within each region may help in locating geographically specific haplotypes or polymorphic sites. Additionally, it would be beneficial to increase the representation of members of the lower castes in order to be better able to detect any intercaste variation. Finally, it would be of interest to obtain tribal samples, as
we may be able to determine more about the history of the original inhabitants of the subcontinent in this manner.
Conclusion

It has been said, regarding the inhabitants of the vast region encompassing the subcontinent of India, that "this people has no history, or at least it has had no historians" (Masson-Oursel et al., 1967). In this study, we attempted to reconstruct some of the complex historical and social processes which have helped in shaping the genetic signature of the peoples of India.

On the basis of sequence analysis of the 195 bp region of the non-coding mitochondrial D-loop sequenced in this study, no major structuring with respect to geography, language or caste was observed. The phylogenetic trees inferred from the data did however display a few minor trends. Speakers of Dravidian languages were significantly under-represented within one cluster, suggesting a possible language-related effect or an effect related to geographical or historical origin. A similar under-representation of members of one caste (the Vaisyas) was also observed in one cluster. A north-south effect was observed in significant discrepancies in the geographical distributions of several polymorphic sites.

Although sensitive enough to resolve differences between major human populations which are genetically more diverged, the technique applied in this study may not be sensitive enough to resolve population structuring within a population of individuals as closely related as those of the Indian subcontinent.
As the extreme variability present in the sequenced region creates a great deal of background variation, which may be obscuring information to be gained from the more polymorphic sites, future research may be better served in using techniques addressing only those sites determined to be polymorphic for this population, such as analysis of RFLP markers.
Appendices
Appendix A - Protocols

Protocols

Blood collection

- Collect in 5 ml draw lavender VACUTAINER brand blood collection tubes. Tubes contain 0.05 ml of 15% EDTA (K₃) solution. *Reorder number 6452.*

- Wear gloves at all times when handling potential DNA samples.

- Place 1 ml aliquots of blood (using autoclaved tips) into autoclaved 1.5 ml microcentrifuge tubes.

- Freeze and store blood aliquots at -70°C.
**DNA Extraction - Phenol**

**Materials:**

- Blood samples
- Lysis buffer (1X)
- Saline
- SSTE
- Proteinase K
- Phenol
- Chloroform iso-amyl alcohol
- Autoclaved 1.5 ml eppendorf tubes
- NaAc (2M, pH 5.2)
- EtOH (100%)
- TE

**Whole blood preparation:**

- Use 1 ml whole blood in 1.5 ml microcentrifuge tube.
- Spin 2 minutes @ 8000 rpm.
- Remove approximately 0.5 ml plasma (using 1000 µl pipetter), leaving cellular material in tube.

**Lyse red blood cells:**

- Add 1 ml cold Red Cell Lysis Buffer (1X).
- Mix by inverting tube a few times.
- Allow to stand on ice for 15 minutes.
- Spin 2 minutes @ 8000 rpm and remove top layer of liquid. The liquid being removed should be clear (dark red the first few times). Stop where the liquid becomes cloudy (remove approximately 0.75 ml).
- Resuspend the pellet well and add 1 ml Red Cell Lysis Buffer. Spin and remove plasma.
- Repeat as necessary (3 lyses should be sufficient).
- When a good white blood cell pellet is obtained (without cellular junk), wash with 300 ml saline (spin @ 8000 rpm for 2 minutes, pipette off supernatant).

**Lyse white blood cells:**

- Add 300 µl SSTE and 100 µl proteinase K (20 mg/ml); total volume = 400 µl
- Resuspend pellet (liquid will be globular).
- Incubate @ 55°C for minimum 1 hour (up to overnight - lose DNA if more than 2-3 hours). The brown protein pellet at the bottom of the tube should be dissolved.
- If suspension is still very gelatinous after incubation, add another 100 µl SSTE and 50 µl proteinase K.
**Separate proteins - Phenol:**

* Procedure is performed in fume hood *

- Add equal volume of phenol (350 µl) and mix by inverting.
- Spin for 2 minutes at 8000 rpm.
- Using pipette, remove aqueous phase and transfer to another (autoclaved) 1.5 ml microcentrifuge tube. At this stage take any proteins (white stringy material) as well.
- Repeat once more using 350 µl phenol.
- if aqueous phase is still brown, repeat again using phenol.
- Repeat using a 1:1 mixture of phenol:chloroform iso-amyl alcohol (175 ml of each). Avoid proteins at this point.
- Repeat using 350µl chloroform iso-amyl alcohol. Stay away from bottom phase and proteins.

**Extract DNA:**

- To last tube add 1/10th volume of 2M NaAc (approximately 35 µl) and 2-3 x volume of 100% ethanol (approximately 700 µl). Mix by inverting tube gently a few times.
- Wait until DNA comes out (white strings). When striations are no longer seen between the phases, spin at 14000 rpm for 6 minutes.
- Pour ethanol into sink. Be careful not to lose pellet.
- Add 700 µl ethanol, resuspend pellet and spin at 14000 rpm for 2 minutes.
- Drain ethanol into sink being very careful not to lose pellet. Blot with kimwipes.
- Incubate at 55°C until dry (5-10 minutes).
- Add 50 µl TE (adjust volume depending on size of pellet if necessary) and mix by flicking tube.
- Incubate @ 55°C for 2-3 hours.
- Store at 4°C (refrigerator) or freeze at 20°C.
DNA Extraction - Salt

Materials:
Sterile 6 ml plastic tubes
Autoclaved 1.5 ml microcentrifuge tubes
1X Red Cell Lysis buffer
Saline
SSTE
Proteinase K (20 mg/ml)

Whole blood preparation:

• Use 1 ml whole blood in 1.5 ml microcentrifuge tube.
• Spin 2 minutes @ 8000 rpm.
• Remove approximately 0.5 ml plasma (using 1000 µl pipetter), leaving cellular material in tube.

Lyse red blood cells:

• Add 1 ml cold Red Cell Lysis Buffer (1X).
• Mix by inverting tube a few times.
• Allow to stand on ice for 15 minutes.
• Spin 2 minutes @ 8000 rpm and remove top layer of liquid. The liquid being removed should be clear (dark red the first few times). Stop where the liquid becomes cloudy (remove approximately 0.75 ml).
• Resuspend the pellet well and add 1 ml Red Cell Lysis Buffer. Spin and remove plasma.
• Repeat as necessary (3 lyases should be sufficient).
• When a good white blood cell pellet is obtained (without cellular junk), wash with 300 ml saline (spin @ 8000 rpm for 2 minutes, pipette off supernatant).

Lyse white blood cells:

• Add 300 µl SSTE and 100 µl proteinase K (20 mg/ml): total volume = 400 µl
• Resuspend pellet (liquid will be globular).
• Incubate @ 55°C for minimum 1 hour (up to overnight - lose DNA if more than 2-3 hours). The brown protein pellet at the bottom of the tube should be dissolved.
• If suspension is still very gelatinous after incubation, add another 100 µl SSTE and 50 µl proteinase K.
Separate proteins - Salt:

- Add 350 µl 5M NaCl and 350 µl DDW (total volume = 1100 µl).
- Mix well (to precipitate protein).
- Centrifuge for 30 minutes @ 4000 rpm.
- Transfer supernatant to a 6 ml sterile plastic tube by pouring. There should be a precipitated brown pellet at the bottom of the tube (protein), and some brown scummy looking material floating near the top (impure DNA). Pour everything except the protein pellet into the 6 ml tube.

Extract DNA:

- Add approximately 100 µl of 2M NaAc (1/10th volume) and approximately 2.5 ml of room temperature 100% ethanol (2-3 x). Mix by inverting tube gently a few times.
- Wait until DNA comes out (white strings).
- If the DNA is still very brown allow it to sit in the EtOH/NaAc solution for a few hours (overnight is fine). The DNA should become more white.
- When striations are no longer seen between the phases, and the DNA is clean, spool out DNA with a heat sealed pasteur pipette.
- Rinse (twice) with 70% ethanol. Pour drop on pipette and allow it to run down to end of pipette over pellet or swirl in fresh 1.5 ml microcentrifuge tube filled with 70% ethanol.
- Allow DNA to air dry for 5 - 10 minutes.
- Resuspend in 50 µl TE (adjust volume depending on size of pellet if necessary) and mix.
- Incubate @ 55°C for 2-3 hours.
- Store @ 4°C (refrigerator) or freeze @ 20°C.


**PCR Amplification of 1 kb region of mtDNA**

1. Heat template DNA samples for @ 95°C for 5'.

2. Label *GeneAmp™* PCR tubes with sample #. Label one extra tube for a negative control.

3. Make Master Mix in negative control tube.

   For each sample (including negative control) add:

<table>
<thead>
<tr>
<th>Volume (μl)</th>
<th>Reagent</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.75</td>
<td>Distilled Deionized Water</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10X PCR Buffer II</td>
<td>(100mM Tris-HCl, pH 8.3 @ 25°C; 500 mM KCl)</td>
</tr>
<tr>
<td>3</td>
<td>MgCl₂ Solution</td>
<td>(25 mM MgCl₂)</td>
</tr>
<tr>
<td>1</td>
<td>dATP</td>
<td>(10 mM)</td>
</tr>
<tr>
<td>1</td>
<td>dCTP</td>
<td>(10 mM)</td>
</tr>
<tr>
<td>1</td>
<td>dGTP</td>
<td>(10 mM)</td>
</tr>
<tr>
<td>1</td>
<td>dTTP</td>
<td>(10 mM)</td>
</tr>
<tr>
<td>1</td>
<td>Primer 1</td>
<td>(L15996)</td>
</tr>
<tr>
<td>1</td>
<td>Primer 2</td>
<td>(H408)</td>
</tr>
</tbody>
</table>

4. Cap tube and mix thoroughly by flicking side of tube.

5. Aliquot 48.75 μl into each sample tube and add:

<table>
<thead>
<tr>
<th>Volume (μl)</th>
<th>Reagent</th>
<th>Final Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>AmpliTaq® DNA Polymerase</td>
<td>(5 Units / μl)</td>
</tr>
<tr>
<td>1</td>
<td>Template DNA</td>
<td></td>
</tr>
</tbody>
</table>

   **Total Volume:** 50 μl

6. To each tube add 25 μl Mineral Oil (1 drop).

7. Cap tubes tightly and mix thoroughly.

8. Spin tubes briefly @ 13000 rpm to gather liquid at bottom.

9. Place in PCR machine and cycle:

   1 cycle: 1' @ 95°C

   30 cycles of:

<table>
<thead>
<tr>
<th>Step</th>
<th>Time (s)</th>
<th>Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denature</td>
<td>45</td>
<td>93</td>
</tr>
<tr>
<td>Anneal</td>
<td>1</td>
<td>55</td>
</tr>
<tr>
<td>Extend</td>
<td>3</td>
<td>74</td>
</tr>
</tbody>
</table>

   Link to 4°C soak file.

10. Visualize on agarose gel or freeze @ -20°C.
GlassMAX™ purification of PCR product

- Begin incubating 50 μl TE per sample @ 65°C.

- Add 4.5 Volumes of Binding Solution (Nal) to each sample;
  100 μl reaction: add 450 μl Nal
  50 μl reaction: add 225 μl Nal.

- Invert to mix.

- Transfer solution to a spin cartridge (max 550 μl).

- Spin @ 13 000 x g for 20 seconds.

- Pour out contents of tube.

- Add 400 μl cold wash buffer to the spin cartridge.

- Spin @ 13 000 x g for 20 seconds.

- Pour out contents of tube.

- Repeat twice.

- After final rinse, spin for 1 minute @ 13 000 g.

- Transfer spin cartridge insert into Sample recovery tube.

- Add 40 μl 65°C TE.

- Spin @ 13 000 x g for 20 seconds.

- Remove spin cartridge and discard.
**Ethanol purification of PCR product**

For a 50 μl reaction:

- Run 5 μl on gel to visualize.
- Take 43 μl from reaction tube and place in 1.5 ml microcentrifuge tube
  
  Add:

  - Add 18 μl of 5M NH₄Ac and 90 μl 95% EtOH (total volume = 43 + 18 + 90 = 151 μl).
  - Mix well by inverting tube.
  - Keep on ice for 10 minutes. (leaving at room temp gives less impurities, but also less precipitation).
  - Spin at 14 000 rpm for 20 minutes.
  - Draw off 135 μl supernatant with Pipetman, leaving 16 μl at bottom (the precipitate is usually not visible).
  - Wash with 200 μl 70% EtOH.
  - Resuspend pellet.
  - Spin @ 14 000 rpm for 10 minutes.
  - Pipette off 205 μl EtOH (< 10 μl at bottom).
  - Dry 5-10 minutes.
  - Resuspend in 50 μl TE.
Cycle Sequencing

Label Primer(s) - $^{33}P$

[1] Into GeneAmp™ PCR Tube place:

<table>
<thead>
<tr>
<th>for 6 samples</th>
<th>for 12 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0 µl</td>
<td>6.0 µl</td>
</tr>
<tr>
<td>1.0 µl</td>
<td>2.0 µl</td>
</tr>
<tr>
<td>0.6 µl</td>
<td>1.2 µl</td>
</tr>
<tr>
<td>0.6 µl</td>
<td>1.2 µl</td>
</tr>
</tbody>
</table>

T4 Polynucleotide Kinase (10 Units/µl)
Primer (10 µM)
10X Kinase Buffer
DDW

In radioactive area add:

1.0 µl 2.0 µl γ$^{[33}P$]-ATP (>5000 Ci/mmole)

6.2 µl 12.4 µl Total Reaction Volume


File #: 12
Incubate @ 37°C for 15 minutes and @ 90°C for 5 minutes
Link to 4°C soak file.

Store for up to 5 days @ -20°C.
Cycling Reaction Procedure

[A] (1) Label 40 *GeneAmp™* PCR tubes with sample # (1-10) in the following colours:

- A: Blue
- C: Red
- G: Black
- T: Green

(2) To each tube add 15 µl Mineral Oil.

(3) For each sample, dispense 2 µl of each of the A, C, G, & T Termination Mixes into the tubes.

[B] (1) For each sample, prepare 30 µl Reaction Mix:

<table>
<thead>
<tr>
<th>Volume</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 µl</td>
<td>10X Cycling Mix</td>
</tr>
<tr>
<td>25-x µl</td>
<td>Autoclaved DDW</td>
</tr>
<tr>
<td>x µl</td>
<td>Template DNA (10-100 fmoles)</td>
</tr>
</tbody>
</table>

(2) In radioactive area add:

<table>
<thead>
<tr>
<th>Volume</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 µl</td>
<td>[33P]-End-labelled Primer (1.6 pmoles/µl)</td>
</tr>
<tr>
<td>30 µl</td>
<td>Total volume per sample</td>
</tr>
</tbody>
</table>

(3) Mix.

[C] (1) To each tube prepared in part 1, add:

<table>
<thead>
<tr>
<th>Volume</th>
<th>Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 µl</td>
<td>of the appropriate Reaction Mix</td>
</tr>
</tbody>
</table>

(2) Mix.

(3) Spin tubes to gather liquid at bottom.

[D] (1) Place in PCR machine and cycle:

1 cycle: 1' @ 95°C

20 cycles of:

<table>
<thead>
<tr>
<th>Primer</th>
<th>File #</th>
<th>Denature</th>
<th>Extend</th>
<th>Program Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>OT:</td>
<td>25</td>
<td>95°C</td>
<td>55°C</td>
<td>70 minutes</td>
</tr>
<tr>
<td>IB:</td>
<td>27</td>
<td>95°C</td>
<td>65°C</td>
<td>70 minutes</td>
</tr>
</tbody>
</table>

Link to 4°C soak file.

[E] (1) Add 4 µl Stop Solution to all tubes. Total volume / tube = 12 µl.

(2) Spin tubes to gather liquid at bottom.

Store for up to 1 week @ -20°C.
Preparing Acrylamide Gel:

To make an 8% denaturing Gel:

# gels:

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2-3)</th>
<th>(3-5)</th>
<th>(4-6)</th>
<th>(5-8)</th>
<th>Reagent</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>7.6 g</td>
<td>15.2 g</td>
<td>22.8 g</td>
<td>30.4 g</td>
<td>38.0 g</td>
<td>Acrylamide</td>
</tr>
<tr>
<td>g</td>
<td>0.4 g</td>
<td>0.8 g</td>
<td>1.2 g</td>
<td>1.6 g</td>
<td>2.0 g</td>
<td>Bis-acrylamide</td>
</tr>
<tr>
<td>g</td>
<td>42.0 g</td>
<td>84.0 g</td>
<td>126.0 g</td>
<td>168.0 g</td>
<td>210.0 g</td>
<td>Urea</td>
</tr>
<tr>
<td>ml</td>
<td>10 ml</td>
<td>20 ml</td>
<td>30 ml</td>
<td>40 ml</td>
<td>50.0 ml</td>
<td>10X TBE buffer</td>
</tr>
<tr>
<td>ml</td>
<td>40 ml</td>
<td>80 ml</td>
<td>120 ml</td>
<td>160 ml</td>
<td>200.0 ml</td>
<td>DDW</td>
</tr>
</tbody>
</table>

100 ml 200 ml 300 ml 400 ml 500 ml Total Volume

- Combine in erlenmeyer flask. Heat and stir until solids are completely dissolved (@ approximately 60°C).
- Filter through filter mechanism into 250/500 ml erlenmeyer flask.
- Adjust volume to 100/200/300/400/500 ml with DDW.
- De-gas for 2 minutes by swirling under vacuum.
- Store in brown bottle or in erlenmeyer in fridge.
To make mould:

- Make up 900 ml of 1X TBE buffer in 1 L Erlenmeyer flask.
- Rinse plates with DDW and dry with Kimwipe. Rinse with Ethanol and dry with Kimwipe.
- Siliconize short plate only: In fumehood pour small amount of Sigmacote onto plates and wipe over plate with Kimwipe. Allow to air dry for at least 10 minutes.
- Place spacers between plates. Tape glass plates together. The unnotched (clean) sides go inwards. With short plate on top, notched corners are to the right.
- Rinse sharkstooth combs with Ethanol.
- Place two 2.5" bulldog clamps along each side near top using bulldog clips. Make sure clamps are positioned above spacers and no further inwards. If using double-fine combs, place one 2.5" bulldog clamp over top in middle.

To pour gel:

- To 60/70/75/100 ml of 8% Acrylamide solution add:

  300/350/375/500 μl 10% APS
  
  • swirl

  30/35/37.5/50 μl TEMED
  
  • swirl

- Pour gel using 60 ml syringe (no needle). Hold mould at 45° angle and pour acrylamide along one edge of plate.
- Push backs of sharkstooth combs into gel (up to edge of holes in combs). Cover with acrylamide.
- Lay the top of the mould on an object which will allow the gel to polymerize at a 10° angle.
- Place the 1 L erlenmeyer flask filled with 1 L of 1X TBE on gel.
- When gel is completely polymerized, cover top with 1X TBE soaked paper towels and plastic wrap.
Running Gel: short run

- Wash off extra acrylamide from outside of plates and rinse with DDW.
- Remove clips and tape from gel mould.
- Attach plates to apparatus (long plate in front). Tighten stops so that a good seal is formed at top, but do not overtighten.
- Fill top and bottom reservoirs with 450 ml each of 1X TBE (make sure tap is closed). Buffer level should be above bottom of glass plates and above top of short plate.
- Rinse top of gel with buffer (using syringe with needle) to remove excess urea.
- Push combs most of the way into the wells making sure teeth do not touch the top of the gel.
- Preheat @ 2000 V / 65 mA / 50 W for 30 minutes until gel temperature is 55°C - 65°C.
- After 25 minutes, start incubating @ 76°C for 5 minutes (File # 9). Set alarm for 5 minutes.
- Turn off power supply @ 30 minutes.
- Take out both combs. Rinse top of gel to remove excess urea.
- Re-insert combs positioning so that teeth touch the top of the gel.
- Load short run: 1.8 - 2 µl for double fine combs, 2.8 µl for regular combs. Load in order [ A C G T ].
- Approximately 1½ - 2 hours later: When blue dye has partly run off gel, turn off power and unplug.
Running Gel: long run

- Wash off extra acrylamide from outside of plates and rinse with DDW.

- Remove clips and tape from gel mould.

- Attach plates to apparatus (long plate in front). Tighten stops so that a good seal is formed at top, but do not overtighten.

- Fill top and bottom reservoirs with 500 ml each of 1X TBE (make sure tap is closed). Buffer level should be above bottom of glass plates and above top of short plate.

- Rinse top of gel with buffer (using syringe with needle) to remove excess urea.

- Push combs most of the way into the wells making sure teeth do not touch the top of the gel.

- Preheat @ 2000 V / 65 mA / 50 W for 30 minutes until gel temperature is 55°C - 65°C.

- After 25 minutes, start incubating @ 76°C for 5 minutes (File # 9). Set alarm for 5 minutes.

- Turn off power supply @ 30 minutes.

- Take out both combs. Rinse top of gel to remove excess urea.

- Load long run: 2.5 μl for double fine combs, 3.5 μl for regular combs. Load in order [A C C G].

- Approximately 2½ hours later: When cyan dye has reached bottom lower knobs of apparatus, turn off power and unplug.

- If doing a short run, rinse top of gel with buffer (using syringe with needle) to remove excess urea, and load short run. Otherwise, load 2 μl stop solution in outside lane(s). Re-plug in and press start.

- Approximately 1¼ - 2 hours later: When blue dye has partly run off gel, turn off power and unplug.
Removing and drying gel:

- Open tap and allow upper buffer tray to empty.
- Wet a trimmed sheet of Whatman paper with DDW, leaving one inch on either end dry to facilitate handling.
- Remove gel mold from apparatus and place behind radioactive shield with the short plate on top.
- Slide out combs and side spacers. Carefully pry off short glass plate, making sure gel is stuck to bottom and not top plate.
- Place piece of Whatman paper on gel. Gently press on paper. Carefully peel paper (with gel) off the glass and invert so gel is on top.
- Place plastic wrap on top of gel avoiding air bubbles.
- Place gel on dryer with paper on bottom and close dryer.
- Turn on vacuum (water tap) and ensure that a good seal is formed.
- Select drying cycle for sequencing gels: 
  \[ 80^\circ C \]
- Dry for approximately 1 1/4 hours.

Meanwhile...

- Rinse out tubing between chambers of sequencing apparatus by squirting DDW into drain in top chamber. Take out tray and empty. Rinse and return to apparatus. Close tap.
- Clean glass plates (carefully) with detergent and a cloth, then soak in dish pan containing 1M KOH in EtOH. Do not place more than 1 glass plate in KOH solution at a time (unless using thick spacers between plates) or they will become glued to each other.
- Wash plates with soap, then rinse with DDW and then EtOH. Place on cardboard with clean sides inwards.
- In darkroom: Place gel in cassette with gel side up. Place film on top.
- If using glass plates, place yellow piece of paper between glass and film.
146

Miscellaneous:

**Storage:**

- 4 C Refrigerator
- -20 C Freezer
- -70 C Chest Freezers

1X, 10X, 20X are arbitrary measures of stock solution concentrations which have no relation to Molarity.

x % volume is measured either by weight/volume (in g/ml): a 10% solution of a solid in a liquid would be 10 g/100 ml (final volume); or by volume/volume: a 10% solution of a liquid in a liquid would be 10 ml in 90 ml for a total volume of 100 ml.

To dilute X or % solutions to weaker concentrations:

\[
\text{Amount of stock solution (ml)} = \frac{[\text{dilution needed}] \times \text{vol}_T \text{ (ml)}}{[\text{stock on hand}]}
\]

Add appropriate amounts of stock solutions and adjust to vol$_T$ with DDW.
Appendix B - Solutions

Recipes

Stock Solutions

EDTA

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>0.5 M</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td>7.4 &amp; 8.0</td>
</tr>
<tr>
<td>Amount:</td>
<td>1 L</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>Yes</td>
</tr>
<tr>
<td>Storage:</td>
<td>Bench - 500 ml Gibco bottles</td>
</tr>
<tr>
<td>Notes:</td>
<td>Use disodium salt EDTA</td>
</tr>
</tbody>
</table>

\[ \text{M.W.} = 372.24 \]

\[ \begin{align*}
186.12 \text{ g} & \quad \text{disodium salt EDTA} \\
800 \text{ ml} & \quad \text{DDW}
\end{align*} \]

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

HCl

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>1 M</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td></td>
</tr>
<tr>
<td>Amount:</td>
<td>500 ml</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>No</td>
</tr>
<tr>
<td>Storage:</td>
<td>Bench - 500 ml Gibco bottle</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{M.W.} = 36.46 \]

\[ \begin{align*}
18.23 \text{ g} & \quad \text{HCl} \\
400 \text{ ml} & \quad \text{DDW}
\end{align*} \]

NaOH

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>1 M</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td></td>
</tr>
<tr>
<td>Amount:</td>
<td>500 ml</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>No</td>
</tr>
<tr>
<td>Storage:</td>
<td>Bench - 500 ml Gibco bottles</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

\[ \text{M.W.} = 40.00 \]

\[ \begin{align*}
20.00 \text{ g} & \quad \text{NaOH} \\
400 \text{ ml} & \quad \text{DDW} \rightarrow 500 \text{ ml}
\end{align*} \]
NaCl

Concentration: 5M
pH: 
Amount: 500 ml
Autoclave: Yes
Storage: Bench in 500 ml Gibco bottle
Notes: NaCl must be added to the water in increments, ensuring the last addition is completely dissolved.

146.1 g NaCl
400 ml DDW → 500 ml

Concentration: 6M
pH: 
Amount: 100 ml
Autoclave: Yes
Storage: Bench in 100 ml Gibco bottle
Notes: NaCl must be added to the water in small increments, ensuring each last addition is completely dissolved.

35.07 g NaCl
80 ml DDW → 100 ml

SDS

Concentration: 10%
PH: 
Amount: 500 L
Autoclave: Yes
Storage: Bench - 500 ml Gibco bottle
Notes: Wear face mask when handling SDS crystals

50 g SDS
400 ml DDW → 500 ml

Concentration: 20%
PH: 
Amount: 100 ml
Autoclave: Yes
Storage: Bench - 100 ml Gibco bottle
Notes: Wear face mask when handling SDS crystals

20 g SDS
70 ml DDW → 1000 ml

M.W. = n/a
STE

Concentration: 1X  
pH: 7.5 ??  
Amount: 500 ml  
Autoclave: Yes  
Storage: Bench - 500 ml Gibco bottle  
Notes:

1X Stock solution:

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

- adjust pH with NaOH pellets
- adjust volume to 500 ml with DDW

Tris (Tris-HCl)

Concentration: 2 M  
pH: 7.5 & 8.0  
Amount: 1 L / 500 ml  
Autoclave: Yes  
Storage: Bench - 500 ml Gibco bottles  
Notes: Tris must be at room temperature for an accurate pH measurement

242.28 g Tris M.W. = 121.14
800 ml DDW → 1 L

or

141.14 g Tris
400 ml DDW → 500 ml

- adjust pH with HCl (this will take forever)
- adjust volume to 1 L / 500 ml with DDW
Buffers

TAE (Tris Acetate)

| Concentration: | 10X |
| pH:            | 7.8 |
| Amount:        | 1 L |
| Autoclave:     | No  |
| Storage:       | Bench - 500 mL Gibco bottles |

Notes:

- 48.4 g Tris
- 11.4 ml Glacial Acetic Acid
- 20 ml 0.5 M EDTA @ pH 8.0
- 700 ml DDW

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

TBE (Tris Borate)

| Concentration: | 10X |
| pH:            | 8.2 |
| Amount:        | 1 L |
| Autoclave:     | No  |
| Storage:       | Bench - 1 L Brown glass bottle |

Notes: TBE will crystallize as it gets old. Any solutions with crystallization should be discarded.

<108 g Tris
- 55 g Boric Acid
- 40 ml 0.5 M EDTA @ pH 8.0
- 700 ml DDW → 1 L
**DNA Extractions**

**Red Cell Lysis Buffer (NH\textsubscript{4}Cl  Na\textsubscript{4}HCO\textsubscript{3})**

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>10X</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td></td>
</tr>
<tr>
<td>Amount:</td>
<td>1 L</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>Yes</td>
</tr>
<tr>
<td>Storage:</td>
<td>Bench - 500 ml Gibco bottle</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>1X</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td></td>
</tr>
<tr>
<td>Amount:</td>
<td>100 ml</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>Yes</td>
</tr>
<tr>
<td>Storage:</td>
<td>4°C in 100 ml Gibco bottle</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
70.00 \text{ g} & \quad \text{NH}_4\text{Cl} & 1.31 \text{ M} & \text{M.W.} = 53.45 \\
0.71 \text{ g} & \quad \text{NH}_4\text{HCO}_3 & 0.009 \text{ M} & \text{M.W.} = 78.98 \\
800 \text{ ml} & \quad \text{DDW} \rightarrow 1 \text{ L} &             &             \\
\end{align*}
\]

**Saline**

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>0.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td></td>
</tr>
<tr>
<td>Amount:</td>
<td>500 ml</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>Yes</td>
</tr>
<tr>
<td>Storage:</td>
<td>Bench - 100 ml Gibco bottles</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
4.5 \text{ g} & \quad \text{NaCl} &             &             \\
400 \text{ ml} & \quad \text{DDW} \rightarrow 500 \text{ ml} &             &             \\
\end{align*}
\]

**SSTE (white blood cell lysis buffer)**

<table>
<thead>
<tr>
<th>Concentration:</th>
<th>0.5% SDS in STE</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH:</td>
<td></td>
</tr>
<tr>
<td>Amount:</td>
<td>100 ml</td>
</tr>
<tr>
<td>Autoclave:</td>
<td>Yes</td>
</tr>
<tr>
<td>Storage:</td>
<td>Bench - 100 ml Gibco bottle</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
5 \text{ ml} & \quad 10\% \text{ SDS} \\
95 \text{ ml} & \quad \text{STE} \\
\end{align*}
\]
NaAc

Concentration: 2 M
pH: 80 ml
Amount: NaAc
Autoclave: 27.22 g
Storage: Bench 100 ml Gibco bottle
Notes: M.W. = 136.08

2.5 ml Tris @ pH 7.5 10 mM
1.0 ml EDTA @ pH 8.0 1 mM
450 ml DDW → 500 ml

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

TE

Concentration: 1X
pH: 7.7
Amount: 500 ml
Autoclave: Yes
Storage: Bench - 100 ml Gibco bottles
Notes:

Concentration: 1X
pH: 8.0
Amount: 500 ml
Autoclave: Yes
Storage: Bench - 100 ml Gibco bottles
Notes:

2.5 ml Tris @ pH 8.0 10 mM
1.0 ml EDTA @ pH 8.0 1 mM
450 ml DDW → 500 ml

• adjust pH with HCl
• adjust volume to 500 ml with DDW
### Proteinase K

<table>
<thead>
<tr>
<th><strong>Concentration:</strong></th>
<th>10 mg/ml DDW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pH:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Amount:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Autoclave:</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>Storage:</strong></td>
<td>-20°C in 1.5 ml microcentrifuge tubes</td>
</tr>
<tr>
<td><strong>Notes:</strong></td>
<td>Boehringer-Mannheim (# 745 723). Purchase in 100 mg bottles.</td>
</tr>
</tbody>
</table>

Add 10 ml DDW to 100 mg Proteinase K. Aliquot into autoclaved 1.5 ml Microcentrifuge tubes.

### Ethanol (EtOH)

<table>
<thead>
<tr>
<th><strong>Concentration:</strong></th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pH:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Amount:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Autoclave:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Storage:</strong></td>
<td>Bench in Plastic bottle</td>
</tr>
<tr>
<td><strong>Notes:</strong></td>
<td>Purchase at Scientific Stores</td>
</tr>
</tbody>
</table>
### Reagents and Kits

#### PCR Reagent Kit

<table>
<thead>
<tr>
<th>Product Name:</th>
<th>GeneAmp® PCR Reagent Kit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer:</td>
<td>Perkin Elmer Cetus</td>
</tr>
<tr>
<td>Part Number:</td>
<td>N801-0055</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

#### PCR Product Purification Kit

<table>
<thead>
<tr>
<th>Product Name:</th>
<th>GLASSMAX®</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer:</td>
<td>Gibco BRL (Life Technologies Inc)</td>
</tr>
<tr>
<td>Part Number:</td>
<td>15590-052</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

#### 100 bp ladder

<table>
<thead>
<tr>
<th>Product Name:</th>
<th>100 bp DNA ladder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer:</td>
<td>Gibco BRL (Life Technologies Inc)</td>
</tr>
<tr>
<td>Part Number:</td>
<td>15628-019</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

#### Sequencing Reagent Kit

<table>
<thead>
<tr>
<th>Product Name:</th>
<th>AmpliCycle™ Sequencing Kit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer:</td>
<td>Perkin Elmer Cetus</td>
</tr>
<tr>
<td>Part Number:</td>
<td>N808-0175</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>

#### Sequencing Gels

<table>
<thead>
<tr>
<th>Product Name:</th>
<th>Sequagel-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturer:</td>
<td>National Diagnostics</td>
</tr>
<tr>
<td>Part Number:</td>
<td>EC-838</td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
</tr>
</tbody>
</table>
Appendix C - Algorithms

DECLARE SUB comparesequences()
DECLARE SUB inputsequences()
REM compute pairwise differences between sequences
COMMON SHARED numseqs, numbps, seqfile$, outfile$, qpoutfile$
DIM SHARED sequence$(500)
CALL inputsequences
CALL comparesequences
PRINT
PRINT "Finished!"
PRINT
PRINT "ASCII output saved in file "; outfile$
PRINT "QPro output saved in file "; qpoutfile$
PRINT
END

SUB comparesequences
REM compare sequences for pairwise differences
outfile$ = seqfile$ + ".out"
qpoutfile$ = seqfile$ + ".qp"
OPEN outfile$ FOR OUTPUT AS 2
OPEN qpoutfile$ FOR OUTPUT AS 3
FOR n = 1 TO numseqs
    FOR m = n + 1 TO numseqs
        diffs = 0
        diffs$ = ""
        FOR p = 1 TO numbps
            IF MID$(sequence$(n), p, 1) <> MID$(sequence$(m), p, 1) THEN
                diffs = diffs + 1
                diffs$ = diffs$ + STR$(p)
                'PRINT "difference at:"; p
            END IF
        NEXT p
        PRINT #2, n, m, diffs, diffs$
        WRITE #3, n, m, diffs
    NEXT m
NEXT n
CLOSE 2
CLOSE 3
END SUB
SUB inputsequences
REM input sequences into array
DIM tempseq$(4)
PRINT ""
INPUT "Enter sequence filename:"; seqfile$
OPEN seqfile$ FOR INPUT AS 1
INPUT #1, numseqs, numbps

rows = 0
IF numbps = 195 THEN rows = 3
IF numbps > 195 THEN rows = 4
REM troubleshooting
IF numbps < 195 THEN
PRINT "Mayday!!!"
SLEEP
END IF
numlastline = numbps - 195
FOR s = 1 TO numseqs
LINE INPUT #1, sampleid$
FOR r = 1 TO 3
    tempseq$(r) = INPUT$(65, #1)
    LINE INPUT #1, endcharacters$
NEXT r
    tempseq$(4) = ""
IF rows = 4 THEN
    tempseq$(4) = INPUT$(numlastline, #1)
    LINE INPUT #1, endcharacters$
END IF
sequence$(s) = tempseq$(1) + tempseq$(2) + tempseq$(3) + tempseq$(4)
IF LEN(sequence$(s)) <> numbps THEN
    PRINT "ERROR!!!"
    SLEEP
END IF
NEXT s
CLOSE 1
END SUB
References


