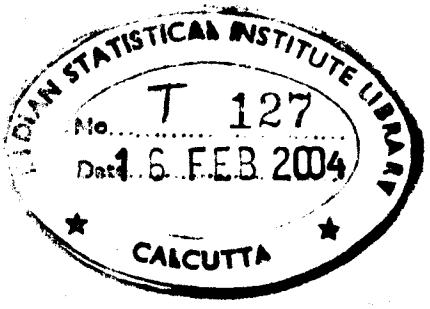
# SOME STATISTICAL CONTRIBUTIONS TO THE ANALYSIS OF HUMAN GENOME DIVERSITY AND EVOLUTION

**Analabha Basu**
Anthropology & Human Genetics Unit
Indian Statistical Institute
Kolkata 700108, India

2003




*A thesis submitted in partial fulfillment of the degree of Doctor of Philosophy of the Indian Statistical Institute*

# ACKNOWLEDGEMENTS

Writing the acknowledgements for this thesis is turning out to be an almost surreal experience to me. As I sit down to write, faces come flocking in. Some of them are almost tangible, some are hard to recognize. Many people contributed, each in their own way, during my tenure as a research fellow. A due reference to all of them will make this thesis twice its present size. I shall therefore start with my sincerest apologies to all those whom I may miss completely as well as to those whom I may fail to acknowledge with due gratitude.

My vocabulary is incomplete and expressions are inadequate to rightly appropriate my gratitude towards Professor Partha Pratim Majumder. Ever since he introduced me to the field of population genetics, it has been a wonderful learning experience with him. His immaculate insight, strive for precision and ability to work hard inspired me a lot. But more than anything else, I am grateful to him as an individual because of the compassion with which he treated me during my difficult days. My years of association with him has definitely made me more matured, both academically and otherwise.

I shall take this opportunity to thank my relatives, especially my parents, who have been very supportive, not only during this tenure but throughout my life.

My special thanks go to the members of our laboratory as well as to those who collaborate with us for doing all the hard work to generate the data that I used in my thesis. I shall take this opportunity to thank all the members of different units of my Institute, particularly the Dean's Office and the Library Unit, for the help they have been to me during various phases of my research. Besides helping academically, it was a pleasure to work with Dr. Bidyut Roy, Mr. Badal Dey, Ms. Monami Roy, Mr. Madan Chakraborty, Dr. Sanghamitra Sengupta, Dr. Namita Mukherjee, Ms. Sangita Roy, Mr. Nilabja Sikdar, Mr Arindam Das Adhikari and various other students and teachers who visited our laboratory.

The fact that I cherished my stay in the institute is largely due to my friends. We really had each other during our best as well as the worst of times. I am indebted to Abhimanyu, Anil, Anirban, Anusthup, Debrup, Partha, Rachana, Subhadip, Sabyasachi, Sasthi, Tapas and many others for making my tenure so memorable.

I gained enormously from the long discussions I had with Professor Probal Chaudhuri and Professor Richard Hudson. I am thankful for the patience they showed in answering my innumerable queries. Professor Hudson also introduced me to his versatile coalescent program that has been so popular worldwide.

I am thankful to Professor Arnold Neumaier for troubleshooting my problems in handling his quadratic optimization program MINQ. I am also thankful to the anonymous reviewers of Human Genetics, Journal of Genetics and Genome Research whose comments helped to improve the contents and clarity of the thesis. I wish to convey my

special thanks to Mr. Rabindra Narayan Das, who took the pain of typing a portion of this thesis.

It will be criminal not to mention Dr. Saurabh Ghosh and Mr. Sujit Maiti. While Sujit tried hard to teach me the intricacies of efficient handling of computers, Saurabh tried to improve my efficiency in general. Though both failed to accomplish their goals, I am thankful to them for their sincere efforts.

Lastly, I thank my best friend and wife Swati, who has been a constant source of energy and inspiration in the endeavor. It was really her ability and courage to stand up to adversities that made this thesis possible.

To end, I shall revert back to where I started from; faces, or rather the lack of it. I am grateful to those "faceless" individuals who voluntarily donated blood to make this study possible.

Kolkata, August 2003 Analabha Basu

# CONTENTS

# CHAPTER 1

## Introduction, Objectives and Overview

# Introduction

The work embodied in this thesis pertains to human population genetics. In particular, the overarching goals of this thesis are to contribute to the understanding of genomic diversity of human populations and to the development of statistical methods for making inferences in genome diversity studies. With these two goals in mind, we have carried out a detailed statistical analysis of genomic data on a large number of ethnic populations of India, generated in the laboratory of the Anthropology & Human Genetics Unit, Indian Statistical Institute, Kolkata. Additionally, wherever relevant, we have compared our data with those collated from the published literature. During the course of this empirical statistical study (the results of which are presented in *Chapter 2*), several methodological issues arose, which resulted in (a) development of probabilistic search algorithms for identifying motifs from DNA sequence data (*Chapter 3*), (b) comparisons of popular methods for estimating time to most recent common ancestor from DNA sequence data (*Chapter 4*), and, (c) development of a statistical method for estimating relative coalescent times from a sample of DNA sequences (*Chapter 5*).

## Chapter 1: Genomic Diversity in India, with Special Reference to Peopling and Population Structure[1]

Based on results of many earlier studies, it is now acknowledged that India occupies a centerstage in human evolution. India has served as a major corridor for the dispersal of modern humans that started from out-of-Africa about 100,000 years before present (ybp). The date of entry of modern humans into India, however, remains uncertain. Further, the migration routes of modern humans into India continue to remain somewhat enigmatic, and whether there were also returns to Africa from India/Asia remain unclear.

Contemporary ethnic India is a land of enormous genetic, cultural and linguistic diversity. It has been shown that, with the exception of Africa, India harbors more genetic

[1] **No references in support of the statements made in this Chapter are provided, since the references are provided in subsequent Chapters.**

diversity than other comparable global regions. The contemporary people of India are culturally stratified as tribals and non-tribals. It is generally accepted that the tribal people are the original inhabitants of India. There are an estimated 461 tribal communities in India, who speak about 750 dialects which can be classified into one of the following three language families: Austro-Asiatic (AA), Dravidian (DR) and Tibeto-Burman (TB). There is considerable debate about the evolutionary histories of the Indian tribals. The proto-Australoid tribals, who speak dialects belonging to the Austric linguistic group, are believed to be the basic element in the Indian population. Many other anthropologists, historians and linguists have also supported the view that the Austro-Asiatic (a subfamily of the Austric language family) speaking tribals to be the original inhabitants of India. Some other scholars have, however, proposed that the Dravidians are the original inhabitants; the Austro-Asiatics are later immigrants. It is, however, noteworthy that the Indian Austro-Asiatic speakers are exclusively tribal, which may be indicative of their being the oldest inhabitants of India. Some believe that the Austro-Asiatic linguistic family evolved in southern China. If indeed this is true, then Indian Austro-Asiatic speakers must have entered India from southern China through the northeast. Many linguists contend that Elamo-Dravidian languages may have originated in the Elam province of southwestern Iran, and the dispersal of the Dravidian languages into India took place with migration of humans from this region who brought with them the technologies of agriculture and animal-domestication. The Tibeto-Burman speaking tribals, who primarily inhabit the north-east regions of India, are supposedly immigrants to India from Tibet and Myanmar.

Most contemporary non-tribal populations of India belong to the Hindu religious fold and are hierarchically arranged in four main caste classes, viz. Brahmin (priestly class), Kshatriya (warrior class), Vysya (business class) and Sudra (menial labour class). In addition, there are several religious communities, who practice different religions, viz. Islam, Christianity, Sikhism, Judaism, etc. The non-tribals predominantly speak languages that belong to the Indo-Aryan or Dravidian families. These two linguistic groups have been the major contributors to the development of Indian culture and society.

Indian culture and society are also known to have been affected by multiple waves of migration that took place in historic and prehistoric times. In a recent study conducted on ranked caste populations sampled from one southern Indian State (Andhra Pradesh), it has been found that the genomic affinity to Europeans is proportionate to caste rank, the upper castes being most similar to Europeans, particularly East Europeans. The lower castes were more similar to Asians. Whether this conclusion can be generalized to caste groups resident in other geographical regions of India remains to be investigated.

As evident from the foregoing discussion, there are considerable differences of opinion among anthropologists and linguists regarding the origins of Indian ethnic groups. Some have argued that human evolution has been largely governed by microevolutionary mechanisms. Therefore, it is crucial to investigate geographically and culturally disparate, but ethnically well-defined, populations in order to understand evolutionary mechanisms that have resulted in the peopling of India. It is also important to statistically analyze data jointly on mitochondrial, Y-chromosomal and autosomal markers from the same populations or sets of populations to gain a comprehensive insight into evolutionary mechanisms. Unfortunately, the vast majority of earlier studies on Indian populations have been conducted on ethnically ill-defined populations or have been restricted to a single geographical area or a single set of markers – primarily either mitochondrial or Y-chromosomal. **The objective of the present study** is to provide a comprehensive view of genetic diversity and differentiation in India and to draw inferences on the peopling of India and the origins of the ethnic populations.

We report a comprehensive study of a large number of ethnic populations of India based on mitochondrial, Y-chromosomal and autosomal markers. Our results indicate that the tribal and the caste populations are genetically highly differentiated. The four linguistic groups of tribals present in India, as also the upper castes of different geographical regions, are also highly differentiated. The Austro-Asiatic tribals seem to be the earliest settlers in India, as evidenced by their large nucleotide diversity and high frequencies of some ancient markers. Y-chromosomal haplogroup frequencies and their present geographical distribution indicate that they may have entered India through the

northeast. There is significant sharing of a small number of mtDNA haplotypes across populations indicating that the number of ancestral female lineages in India was small and also that there has been considerable female movement from one population to another. Subsequent immigrations into India appear to have been predominantly of males. The Tibeto-Burmans tribals, who also entered India from the northeast, share genetic commonalities with the Austro-Asiatic tribals but can be differentiated from them on the basis of Y-STRP haplotypes. The Dravidian tribals, who possibly entered India from the Fertile Crescent region, were possibly widespread throughout India, before the arrival of the Indo-European speaking nomads. After entering through the northwest Indian corridor from Central and West Asia. the Indo-Aryans established their linguistic supremacy over a large number of Dravidian tribals and brought many of them under the fold of the Hindu caste system, which they had formed after adopting a settled life. Many Dravidians, tribals and also castes, seem to have retreated to the southern regions possibly to retain their linguistic and other cultural identities. Indo-European trbals, were probably originally Dravidian speakers, but later adopted the Indo-European speech. Formation of populations by fission and cultural practices that evolved with the caste system have left their imprints on the genetic structures of contemporary populations. Historical migrations into India have contributed to a considerable obliteration of genetic histories of contemporary populations so that there is currently no clear congruence of genetic and geographical or socio-cultural affinities.

In arriving at the above conclusions, we have carried out extensive statistical analyses of genomic data, which included parametric and non-parametric tests of significance, analysis of molecular variance, estimation of nucleotide diversity, phylogenetic analysis, DNA sequence alignment and analysis, and statistical estimation and inferences from mismatch distributions. During the conduct of this empirical study, we encountered problems in applying standard tests of independence in contingency tables, many of which were sparse. We have, therefore, devised a bootstrap procedure for carrying out test of independence in a sparse contingency table.

# Chapter 3: Identification of Polymorphic Motifs Using Probabilistic Search Algorithms

Single nucleotide polymorphisms (SNPs) that occur in the human genome at roughly 1 per 2 kb spacing on the average are often phylogenetically associated. Various evolutionary mechanisms, including natural selection, maintain the association of specific variant nucleotides at one or more sites, which may not be contiguous. The search for associated nucleotides at a set of polymorphic positions is of interest in studies of common diseases and in evolutionary genetics. We define a set of nucleotides that occurs at a high frequency at multiple polymorphic DNA sites, not necessarily contiguous, in a group of individuals as a "motif". We note that our definition of a motif differs from the conventional definition, as for example that is used for finding regulatory sequences in promoter regions of genes, in two ways: (a) the sites included in our motif definition are polymorphic and (b) the sites may not be contiguous. In conventional problems, search is made for evolutionary conserved motifs at a contiguous set of nucleotide positions. In case-control studies of common diseases, it is of interest to find such motifs and to test whether there are differences in motif frequencies between cases and controls. Motifs that are found in significantly higher frequencies among cases are associated with the disease under study. If variants in multiple genes are indeed involved in the disease, the sites in such a motif may not be contiguous. Similarly, the discovery of such motifs is important in evolutionary genetics. Indeed such motifs have been used to define subhaplogroups of specific clades (haplogroups) of the human mitochondrial (mt) DNA.

It is theoretically possible to discover polymorphic motifs in a set of $N$ aligned DNA sequences, each of length $L$ nucleotides, by examining frequencies in all possible $k \times k$ tables, $k$=2,3,...,$L$. However, this is computationally infeasible. The **purpose of this chapter** is to propose a set of probabilistic search algorithms that may be used for motif finding under different scenarios, and to evaluate their efficiencies using both synthetic and real data sets.

Consider a data matrix $((a_{ij}))_{N\times L}$, where $a_{ij}$ denotes the nucleotide (A,T,G or C) at the j-th polymorphic site (j=1,2,...,L) for the i-th individual (i=1,2,...,N). The data matrix is generated from aligned DNA sequences of a specific genomic segment of N individuals, from which all monomorphic sites have been removed. Let V={1,2,...,L} denote the set of all L polymorphic sites in the data. We propose a stochastic search method, similar in spirit to Metropolis-Hastings version of simulated annealing. Our objective function, E(S), to be maximized is the "frequency of a string (S) of nucleotides at p out of L sites". Instead of maximizing E(S), we shall consider minimizing a monotonically decreasing function, H(S), of E(S). The algorithm is iterative. We start with a string $S_0$ of length p; that is, a set of p distinct nucleotide sites drawn randomly from the L polymorphic sites. In each iterative step an element (a nucleotide at a specific site) of the string $S_0$ is updated. Hence, after p such steps we get a completely updated string. The procedure of updating $S_0$ to $S_1$ is called a *sweep*. Thus, *a sweep comprises p iterative steps*. Let $S_t$ denote the updated string after t sweeps.

We shall use the following notations:

1. Let $S_t = (x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(p)})$.
2. Let $S_t^{(i)}$ denote a string in the (t+1)-th sweep, whose first i ($0 \le i \le p-1$) sites have already been modified.
3. Let $S_t^{(i)}(y)$ denote a string in the (t+1)-th iteration, whose first i ($0 \le i \le p-1$) sites have already been modified and the (i+1)-th site is replaced by site y.
4. Let $H_t^{(i)}$ denote the minimum value that has been found for H(S) in the course of all the necessary evaluations of H(S) till the completion of the i-th iterative step in the (t+1)-th sweep.
5. Let $M_t^{(i)}$ denote the string corresponding to $H_t^{(i)}$.

We initially set $H_0^{(0)}=0$ and $M_0^{(0)}$ as a "null" string. The updating procedure for the i-th element in the (t+1)-th sweep uses the idea underlying the Metropolis-Hastings algorithm which can be described as follows:

We first calculate $\beta_t = c.\ ln(t+1)$; where c is a constant > 0. One site (x) is selected at random from the set $V \backslash S_t^{(i-1)}$; that is, from the set V={1,2,...,L} from which the sites included in the set $S_t^{(i-1)}$ have been removed. We then probabilistically update $x_t^{(i)}$ to $x_{t+1}^{(i)}$ according the following rule:

$$x_{t+1}^{(i)} = \begin{cases} x & \text{with probability } \min(\Lambda,1) \\ x_t^{(i)} & \text{with probability } 1 - \min(\Lambda,1) \end{cases}$$

where, $\Lambda = \exp[-\beta_t \{H(S_t^{(i-1)}(x)) - H(S_t^{(i-1)}(x_t^{(i)}))\}]$.

Obviously, the transition probability from one string to another depends only on the outcome of the previous transition (Markov Property). As is easily understood from the above updation rule, at any step of the iteration, although a string that yields a smaller value of H(S) is accepted with a high probability, to avoid being trapped at a local minimum, the current string with higher value of H(S) may also be retained with a small probability (that crucially depends on the control parameter c).

After each iteration we compare $H(S_t^{(i-1)}(x))$ with $H_t^{(i-1)}$. If, $H(S_t^{(i-1)}(x)) < H_t^{(i-1)}$ then, $H(S_t^{(i-1)}(x))$ is the new value for $H_t^{(i-1)}$ and $M_t^{(i)}$ is the updated string $S_t^{(i-1)}(x)$. Otherwise, we do not change $H_t^{(i-1)}$ and $M_t^{(i)}$. In each iteration, therefore we compare the value of the objective function with the smallest value it has attained thus far. This introduces the concept of elitism, which is popular in evolutionary computation, in our algorithm and is done to avoid being trapped at a local minimum.

The possible stopping rules for terminating sweeps in our algorithm can be:

(i) stop if an upper bound, usually a large preset number dependent on availiability of computing resources, on the total number of sweeps (including new initials, if any) is reached, and

(ii) check if the minimum value of H(S) attained thus far during the algorithm has remained unchanged for a certain (preset) number of sweeps. If so, terminate.

The above algorithm pertains to the situation when the motif length is known. However, in reality, the motif length may not be known. We have proposed a modification of this basic algorithm when the motif length is unknown. We have also devised a statistical test procedure that enables determination of the "best" motif length. We have extensively tested the efficiencies of the proposed algorithms using both synthetic and real data sets, and have determined that the algorithms perform very well even under difficult scenarios.

## Chapter 4: Estimating TMRCA from a Sample of DNA Sequences: A Comparison of Two Popular Statistical Methods

The assumption that underlies the statistical reconstruction of the evolutionary history of a set of contemporary populations is that new populations evolve over time by binary fission from ancestral populations. Looking backwards in time, therefore, a set of contemporary populations will coalesce pairwise at different points of time, until finally there is a coalescent event to the most recent common ancestor (MRCA) of all the populations. Such reconstruction can be done by using DNA sequence data generated from samples of individuals drawn from each of the contemporary populations under consideration. The two major features and parameters to be estimated from such data are (a) the topology of the coalescence events, and (b) the times of coalescence to common ancestors of the populations, including the time to MRCA (TMRCA). Both these and parameters are known to be affected by demographic scenarios that prevailed during the process of evolution.

Although there are several methods available for estimating TMRCA from a sample of DNA sequences, two methods are widely used primarily because of conceptual simplicity and ease of interpretation.

Under the infinite sites model, all information in two DNA sequences is captured by the total number of segregating sites ($S_2$). Since $E(S_2|T_2) = \theta T_2$, one approach of estimating $T_2$, which for a sample of two sequences is the TMRCA, is based on $S_2/\theta$. This and similar approaches are not capable of utilizing prior historical demographic information.

Using Bayes' Theorem, it was noted that if $S_2 = k$ then the distribution of $T_2$ is Gamma with parameters $1+k$ and $1+\theta$. In particular,

$$E(T_2|S_2=k) = (1+k)/(1+\theta)$$

$$Var(T_2|S_2=k) = (1+k)/(1+\theta)^2$$

Researchers considered the problem of estimating the TMRCA of n (>2) sequences by extending the analytical results that hold for n=2 and calculated the number of differences between each pair of sequences whose common ancestor is the root of the tree and then averaged these pairwise differences. It was also observed that this value, $\hat{k}$, of k varied little over plausible reconstructed trees. Thus, k was substituted by $\hat{k}$ in the previous equations for $E(T_2|S_2=k)$ and $Var(T_2|S_2=k)$. In a different study, TMRCA was estimated for multiple sequences by substituting the largest value of k among all pairs in the previous equations. This is not a proper approach, because it has been shown that the maximum number of differences between a pair of sequences chosen from this set of n sampled sequences goes to infinity as n goes to infinity. This is true even when $T_n$ is bounded.

A popular alternative to the above procedures of estimating TMRCA, is to use median-joining network analysis. In this analysis, a genealogy of n individuals is considered as an ultrametric tree, in which the lengths of links are scaled to time and each interior node corresponds to a coalescent event. If there are k ($\leq 2n-2$) links of lengths $t_1$, $t_2$, ... ,$t_k$ on a time scale, and if the clade defined by the $i^{th}$ link carries $n_i$ individuals (i= 1,2,...,k) then the coalescent time t can be expressed as

$$t = (n_1t_1 + n_2t_2 + \ldots + n_kt_k)/n.$$

If $\mu$ denotes the mutation rate, expressed as the expected number of (scored) mutations in a sequence segment per time unit, one may associate to the $i^{th}$ link a Poisson distributed random variable $X_i$ with parameter $\mu_i=t_i\mu$. The random variable $X=(n_1X_1+n_2X_2+\ldots+n_kX_k)/n$, has the expected value

$$E(X)=\{(n_1t_1 + n_2t_2 + \ldots + n_kt_k)/n\}\mu=t\mu$$

and variance

$$V(X)=\{(n_1^2t_1+n_2^2t_2+\ldots+n_k^2t_k)/n^2\}\mu,$$

assuming independence of $X_1, X_2, \ldots, X_k$.

The **purpose of the work presented in this Chapter** is to evaluate the performance of these two methods for estimating the coalescent times from DNA sequence data. The data set consisted of nucleotide sequences from homologous segments of DNA sampled from different individuals. The data generated are similar to haploid nucleotide sequences, such as of the mtDNA HVS I (http://www.hvrbase.org)

We have used a forward propagating algorithm to generate simulated DNA sequence data. In this algorithm a nucleotide sequence of a specified length and base composition is created by a multinomial random number generator with cell probabilities equal to the probabilities of the four bases. A completely homogeneous founding population of a given size is then formed by making the appropriate number of copies of the randomly generated nucleotide sequence. The founding population then evolves in accordance with the Wright-Fisher model, i.e. a new generation is formed by sampling from the previous generation with replacement. The numerical size of the succeeding generations is controlled after the founding population is created. In this study we have considered two demographic scenarios: (a) constancy of population size over generations, and (b) exponential growth in size, allowing for variability in the growth parameter over generations. That is, when the size of a new generation is determined, we randomly selected the appropriate number of sequences from the gene pool of the previous generation with replacement. Then, using the assumed value of the mutation rate, we calculated the expected number of mutations per generation, and determined the number of new mutations to be introduced in each generation. If the expected number of mutations per generation is denoted as y, then we randomly chose and mutated [y] or [y]+1 sites, where [y] denotes the largest integer $\leq$ y. Choice between [y] or [y]+1 was made randomly by generating a random number $u$ from the uniform [0,1] distribution, where [y] was chosen if $u$ was less than y–[y]. Suppose there are $N_t$ individuals in generation $t$, each with data on a sequence of L nucleotide sites. To introduce a new mutations in generation t, a site was chosen with probability $1/(N_t \times L)$ and mutated. If $x_i$ is one such observation, then the mutation is introduced at the nucleotide position

$((x_l/L)-[x_l/L])\times L$ of the $[x_l/L]$-th individual. While introducing the mutation, we did not consider any prior information on mutational histories of the site or the individual, thus allowing for parallel, recurrent and back mutations to occur. This process is thus repeated for a stipulated number of generations. The population thus generated was treated as the present population and a random sample of size n was drawn without replacement. This sample of n sequences then was used to estimate the TMRCA of the population. The estimated TMRCA was compared to the actual number of generations used in the simulation.

Since estimates of TMRCA can be affected by various parameters, we have investigated the effects of variation in four crucial parameters. These are:

(1) The number of bases (L) of the nucleotide sequence; we have used two different values of L – 200 and 400.

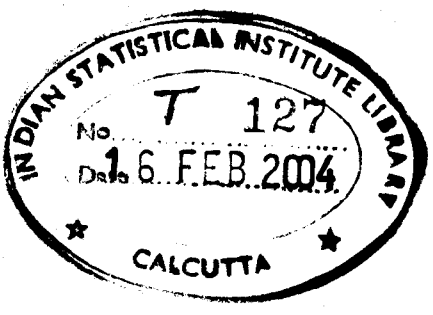
(2) Variability in population size over generation, which was introduced through a parameter $\alpha$. We have used an exponential growth model. In this model, if $N_t$ denotes the population size in generation t, then $N_{t+1} = N_t e^{\alpha}$ . In order that $N_{t+1}$ is an integer, we have chosen either $[N_t e^{\alpha}]$ or $([N_t e^{\alpha}]+1)$. Choice between $[N_t e^{\alpha}]$ or $[N_t e^{\alpha}]+1$ was made randomly by generating a random number $u$ from the uniform [0,1] distribution; $N_{t+1} = [N_t e^{\alpha}]$ was chosen if $u$ was $< (N_t e^{\alpha}) - [N_t e^{\alpha}]$; otherwise $N_{t+1} = [N_t e^{\alpha}]+1$ was chosen. We have used three different values of $\alpha$ – 0, 0.001, 0.005.

(3) The number of generations (g); three different values of g: 250, 500 and 1000 were used.

(4) Mutation rate ($\mu$); two values were used: $10^{-5}$/site/generation and $5\times10^{-5}$/site/generation.

Simulated data were generated using different combinations of the parameter values stated above. For each simulated data set, estimation of TMRCA was carried out using two different methods. TMRCA was estimated from a sample of n=100 sequences.

We have found that the standard deviations (SDs) of the TMRCA estimates were very large, irrespective of the parameter values used in the simulation. Generally, both



methods underestimated the true TMRCA, except for short sequence lengths (L=200, 500) and a short evolutionary time (g=250, 500) with a low mutation rate ($\mu = 10^{-5}$). Both methods were rather insensitive to the population growth parameter ($\alpha$), and there was no consistent trend with respect to $\alpha$ of the either the mean values of the TMRCA estimates or the SDs, although the SDs in many cases decreased with increase in $\alpha$. The frequency distributions of the TMRCA estimates were all highly positively skewed with a very long upper tail for both methods. Our results indicate that in practice considerable caution needs to be exercised in interpreting coalescence times estimated by either of these two methods, which are quite popular.

## Chapter 5: A Statistical Method to Estimate Relative Times of Divergence of Populations from a Common Ancestor

Reconstruction of evolutionary histories of populations is often done from data on DNA sequences from samples of individuals drawn from these populations using phylogenetic methods. The two problems in phylogenetic analysis are (a) estimation of topology, and (b) estimation of branch lengths. It is known from theoretical studies and extensive simulations that correct estimation of topology is easier than estimation of branch lengths with low error. In Chapter 4, we have provided statistical evidence that even the estimate of the time to the most recent common ancestor (TMRCA) can be poor. Additionally, past demographic histories, such as whether the population size has remained constant or whether the population has passed through a bottleneck, affect the phylogenetic relationships among DNA sequences, particularly branch lengths. Even to estimate TMRCA, prior knowledge, or minimally some assumptions, of a population parameter $\theta=4N\mu$ ($\mu$=mutation rate/site/generation) is required, which is often unknown.

The **purpose of this Chapter** is to propose a statistical method to efficiently estimate *relative* branch lengths, which is often sufficient for evolutionary inferences. This method too does not require prior knowledge or make any assumptions on $\theta$.

For a sample of $n$ haploid DNA sequences, let $k_{ij}$ denote the number of mismatches between the *i-th* and *j-th* sequences ($1 \le i < j \le n$); that is, $k_{ij}$ denotes the number of nucleotide positions at which the *i-th* and *j-th* sequences differ. Under the infinite sites model, these $k_{ij}$ differences arose after the two sequences diverged from a common ancestor. If $t_{ij}$ denotes the time since divergence of these two sequences, *i* and *j*, from their common ancestor, and if $\mu$ denotes the mutation rate per site per generation, then

$$E(k_{ij}) = 2\mu\, t_{ij}\,.$$

It can easily be shown that for $n$ sequences,

$$E(\boldsymbol{k}) = 2\,\mu\,\boldsymbol{B}\,\boldsymbol{t}\,,$$

where $\boldsymbol{B}$ is a matrix of zeros and 1s, and $\boldsymbol{k}$ and $\boldsymbol{t}$ have obvious definitions. If $t^* = 2\mu\, t$, then an ordinary least squares estimator of $t^*$ is

$$\hat{\boldsymbol{t}}^* = (\mathrm{B'B})^{-1}\mathrm{B'}k\,,$$

where $\boldsymbol{B}$'= transpose of $\boldsymbol{B}$. The problem with this estimator is that estimates of individual components of $\hat{\boldsymbol{t}}^*$ may be $\le 0$, when in fact times of divergence can not be negative. To ensure that estimates of individual components of $\hat{\boldsymbol{t}}^*$ are $\ge 0$, optimization has to be done on a restricted domain $\boldsymbol{t}^* \ge 0$.

We have used a quadratic programming (QP) approach to carry out optimization on the restricted domain $\boldsymbol{t}^* \ge \mathbf{0}$. The error sums of squares under the model $k = \mathrm{B}t^* + \varepsilon$ is:

$$\begin{aligned}\varepsilon'\,\varepsilon &= (k - \mathrm{B}t^*)'(k - \mathrm{B}t^*) \\ &= k'\,k - 2k'\,B\,t^* + t^{*'}\,(B'\,B)\,t^*.\end{aligned}$$

We need to minimize $\varepsilon'\,\varepsilon$ with respect to $\boldsymbol{t}^* \ge \mathbf{0}$. Now, minimizing $\varepsilon'\,\varepsilon$ is equivalent to minimizing

$$f(t^*) = -2k' B t^* + t^{*'} (B'B) t^*,$$

where $B'B$ is positive semi-definite.

Therefore, $f(t^*)$ is convex, and hence any local minimum is a global minimum of $f(.)$. It is interesting to note that if the ordinary least squares estimate is in the feasible region (i.e., $t^* \geq \mathbf{0}$), then that estimate is the same as the one obtained by minimizing $f(.)$. Our problem, therefore, is to minimize $f(t^*)$, subject to $t^* \geq \mathbf{0}$, which we have carried out using quadratic programming. To obtain estimates of relative times of divergence, which are independent of $\mu$ and hence do not require prior knowledge of $\mu$, we have proposed the natural estimator $\frac{\hat{t}_i^*}{\hat{t}_j^*}$ where $i$=2,3,...,$j$-1 and $j$=3,4,...,$n$. This cancels the multiplicative constant $2\mu$ from both the numerator and the denominator. Under neutrality and constant population size, the expected value of $t_i$ is known. Consider a population that has evolved at constant growth rate (*e.g.*, an exponentially growing population with a growth rate of $\alpha$ per generation). Looking backward in time, the effective population size $t$ generations ago ($N_t$) decreases as $t$ increases. As the probability of occurrence of a coalescent event in generation $t$ is $1/N_t$ , the decrease in $N_t$ with respect to $t$ subsequently increases the probability of a coalescent event in each generation. This phenomenon, therefore, reduces the expected value of $t_j$, which becomes smaller in magnitude as we move backward in time. $E(t_2)$ is, therefore, affected to a greater degree than $E(t_n)$. Consequently, $E(t_2/t_n)$ is much smaller for an exponentially (with parameter $\alpha$) growing ($\alpha$>0) population than a population whose size has remained constant over time. As is evident, the scenario gets reversed when $\alpha$ is < 0, *i.e.*, when we consider a population which has exponentially decreased ($\alpha$<0) in size over time.

A similar pattern is also expected when we consider a population that has passed through a recent bottleneck. The behaviour of coalescence-time, considered as a random variable, in a population that has passed through a bottleneck is similar to a population with a constant effective size until the time of bottleneck. As we look backwards in time, due to the sudden increase in population size prior to the bottleneck, the expected time for

the coalescent events occurring before the bottleneck increases. Thus $E(t_2 \mid \text{bottleneck}) >> E(t_2 \mid \text{constant population size})$. This results in a sudden increase in ratios, such as $E(t_2/t_n)$. These systematic trends, caused by various demographic histories of a population, will be reflected in the estimates of the different ratios of coalescent times.

We have carried out extensive coalescent simulations to assess the performance of this estimator under different demographic history scenarios, and has shown that the estimator behaves as per expectations. We have also proposed appropriate statistical tests that permit demographic-history inferences of populations.

# CHAPTER 2

## Genomic Diversity in India, with Special Reference to Peopling and Population Structure

An abridged version of this Chapter has been accepted for publication in *Genome Research* under the title "Ethnic India: A Genomic View, with Special Reference to Peopling and Structure" authored by Analabha Basu et al.

## Introduction

India occupies a centerstage in human evolution. Several researchers (Nei and Ota 1991; Lahr and Foley 1994; Cavalli-Sforza et al. 1994; Stringer 2000) have proposed that modern humans took two separate dispersal routes from Africa: a 'northern' route through North and East Africa through the Middle East towards Central Asia and western Eurasia and then into India, and a 'southern' coastal route through Ethiopia, Saudi Arabia, Iraq, Iran, Pakistan, along the Indian coastlines, and then further across East Asia to Southeast and South Asia. There is now some evidence that the peopling of the Andaman Islands and Australia took place from India (Endicott et al., 2002; Redd et al. 2002). Thus, India has served as a major corridor for the dispersal of modern humans that started from out-of-Africa about 100,000 years before present (ybp). The date of entry of modern humans into India remains uncertain. However, modern human remains dating back to the late Pleistocene (55000-25000 years before present, ybp) have been found (Kennedy et al. 1987) and by the middle paleolithic period (50,000 – 20,000 ybp), humans appear to have spread to many parts of India (Misra 1992). Kivisild et al. (1999a) have shown that the mtDNA types found in Indian populations belong to the African mtDNA haplogroup L3a, of which 60% belong to the Asian-specific haplogroup M. Further, they have also found that haplogroup U, which was considered to be western Eurasian, is actually the second most frequent haplogroup in India. This haplogroup comprises several subtypes, of which some are found in much higher frequencies in India than in western Eurasia. The coalescence time of the western Eurasian and the Indian U2 subtypes was estimated to be 53,000 ybp. Another haplogroup, U7, found at high frequencies in India and rarely in western Eurasia, has an estimated coalescence time of 32,000 ybp. Thus, it appears, especially because haplogroup U is also present in Ethiopia (Passarino et al. 1998), that diverse north or northeast African gene pool yielded separate origins for western Eurasian and southern Asian populations over 50,000 ybp (Disotell 1999). Genetic support for the southern exit route has also been provided by Quintana-Murci et al (1999). They showed that a specific subclade, M1, of haplogroup M is present in eastern Africa. However, although the four specific transitions characterizing this clade have not been found in the many tribal populations of India examined by Roychoudhury et al. (2001), all these four transitions individually and in various pairwise

combinations were found, indicating that the evidence provided by Quintana-Murci et al. (1999) may be somewhat tentative. More recently, Maca-Meyer et al. (2001) have contended that a posterior return from Asia to Africa of these mtDNA lineages is a more plausible explanation since the genetic diversity of M is much greater in India than in Ethiopia and also because the ancestral motifs of the African M1 are found in M*, M3 and M4 Indian subclusters (Kivisild et al. 1999b). A recent study (Wells et al. 2001) has shown that Central Asia is a land of high genetic diversity and has served as a source of a wave of migration into India, and that the 'diagnostic Indo-Iranian marker' – the M17 polymorphism on the Y chromosome – has a much higher frequency in a Indo-European speaking population of India than in two Dravidian-speaking populations. Thus, the migration routes of modern humans into India continue to remain somewhat enigmatic, and whether there were also returns to Africa from India/Asia (Roychoudhury et al. 2001; Maca-Meyer 2001; Cruciani et al. 2002) remains unclear.

Contemporary ethnic India is a land of enormous genetic, cultural and linguistic diversity. It has been shown that, with the exception of Africa, India harbors more genetic diversity than other comparable global regions (Majumder 1998). The enormous diversity in social and cultural beliefs and practices has been well documented and emphasized (Karve 1961; Beteille 1998). The contemporary people of India are culturally stratified as tribals and non-tribals. It is generally accepted that the tribal people are the original inhabitants of India (Thapar 1966; Ray 1973). The tribals constitute 8.08% of the total population of India (1991 Census of India). There are an estimated 461 tribal communities in India (Singh 1992), who speak about 750 dialects (Kosambi 1991) which can be classified into one of the following three language families: Austro-Asiatic (AA), Dravidian (DR) and Tibeto-Burman (TB). There is considerable debate about the evolutionary histories of the Indian tribals. The proto-Australoid tribals, who speak dialects belonging to the Austric linguistic group, are believed to be the basic element in the Indian population (Thapar 1966, p. 26). Many other anthropologists, historians and linguists (Risley 1915; Rapson 1955; Pattanayak 1998) have also supported the view that the Austro-Asiatic (a subfamily of the Austric language family) speaking tribals to be the original inhabitants of India. Some other scholars (Buxton 1925; Sarkar 1958) have, however, proposed that the Dravidians are the original inhabitants; the Austro-Asiatics

are later immigrants. The Austro-Asiatic family is a fragmented language group. It is most widely spoken in Vietnam and Cambodia. Within India, only a small number of ethnic groups speak Austro-Asiatic languages. It is, however, noteworthy that the Indian Austro-Asiatic speakers are exclusively tribal, which may be indicative of their being the oldest inhabitants of India (Pattanayak 1998; Gadgil et al. 1998). Some believe that the Austro-Asiatic linguistic family evolved in southern China (Diamond 1997). If indeed this is true, then Indian Austro-Asiatic speakers must have entered India from southern China through the northeast. Many linguists (Renfrew 1987; Ruhlen 1991) contend that Elamo-Dravidian languages may have originated in the Elam province of southwestern Iran, and the dispersal of the Dravidian languages into India took place with migration of humans from this region who brought with them the technologies of agriculture and animal-domestication. The Tibeto-Burman speaking tribals, who primarily inhabit the north-east regions of India, are supposedly immigrants to India from Tibet and Myanmar (Guha 1935).

Most contemporary non-tribal populations of India belong to the Hindu religious fold and are hierarchically arranged in four main caste classes, viz. Brahmin (priestly class), Kshatriya (warrior class), Vysya (business class) and Sudra (menial labour class). In addition, there are several religious communities, who practice different religions, viz. Islam, Christianity, Sikhism, Judaism, etc. The non-tribals predominantly speak languages that belong to the Indo-Aryan or Dravidian families. These two linguistic groups have been the major contributors to the development of Indian culture and society (Meenakshi 1995). Indian culture and society are also known to have been affected by multiple waves of migration that took place in historic and prehistoric times (Ratnagar 1995; Thapar 1995). In a recent study conducted on ranked caste populations sampled from one southern Indian State (Andhra Pradesh), Bamshad et al. (2001) have found that the genomic affinity to Europeans is proportionate to caste rank, the upper castes being most similar to Europeans, particularly East Europeans. The lower castes were more similar to Asians. Whether this conclusion can be generalized to caste groups resident in other geographical regions of India remains to be investigated.

As evident from the foregoing discussion, there are considerable differences of opinion among anthropologists and linguists regarding the origins of Indian ethnic

groups. Lahr and Foley (1998) have argued that human evolution has been largely governed by microevolutionary mechanisms. Therefore, it is crucial to investigate geographically and culturally disparate, but ethnically well-defined, populations in order to understand evolutionary mechanisms that have resulted in the peopling of India. It is also important to statistically analyze data jointly on mitochondrial, Y-chromosomal and autosomal markers from the same populations or sets of populations to gain a comprehensive insight into evolutionary mechanisms. Unfortunately, the vast majority of earlier studies on Indian populations have been conducted on ethnically ill-defined populations or have been restricted to a single geographical area or a single set of markers – primarily either mitochondrial or Y-chromosomal. The objective of the present study is to provide a comprehensive view of genetic diversity and differentiation in India and to draw inferences on the peopling of India and the origins of the ethnic populations.

## Materials and Methods

### *Populations*

We have studied a total of 44 populations. The populations were chosen so that they represented ethnic groups of all geographical regions, socio-cultural and linguistic categories. A list of populations is provided in Table 2.1, with brief notes on their socio-cultural backgrounds. The geographical locations of sampling of these populations are indicated in Figure 2.1. It is worth pointing out that (a) population groups of northern India are Indo-European speakers, while those of southern India are Dravidian speakers, (b) the Austro-Asiatic speakers are all tribals and are primarily confined to the central, eastern and northeastern regions, (c) the Tibeto-Burman speakers are confined to the northeastern region, (d) the number of Indo-European speaking tribal groups is very few. Thus, there is an extent of confounding of geography, culture and language in the distribution of ethnic groups of India. This confounding will be reflected in the nature of our statistical analyses and inferences.

Blood samples were drawn from individuals, unrelated to the first cousin level, with informed consent. DNA was isolated from these individuals using Miller et al.'s (1988) protocol. Because of paucity of DNA, some of the populations had to be excluded

Table 2.1
Names of Study Populations, Sample Sizes, Geographical, Linguistic, and Ethnological Information

| POPULATION NAME [CODE] | SAMPLE SIZE | | | | GEOGRAPHICAL DISTRIBUTION | LINGUISTIC AFFILIATION | SOCIAL CATEGORY | OCCUPATION |
|---|---|---|---|---|---|---|---|---|
| | mt | | Y | Auto-somal | | | | |
| | RSP | HVS1 Seq | | | | | | |
| 1. Agharia [AGH] | 24 | 10 | 9 | 24 | Eastern India | Indo-European | Middle Caste | Agriculture |
| 2. Ambalakarer [AMB] | 30 | 10 | 18 | 50 | Southern India - Tamilnadu | Dravidian | Middle Caste | Agricultural labor |
| 3. Bagdi [BAG] | 31 | 10 | 11 | 31 | Eastern India | Indo-European | Lower Caste | Agricultural labor |
| 4. Chakma [CHK] | 10 | 10 | 4 | 10 | North-eastern India - Primarily Tripura | Tibeto-Burman | Tribe | Agriculture |
| 5. Chamar [CHA] | 25 | 10 | 18 | 25 | Northern India | Indo-European | Lower Caste | Menial and agricultural labor |
| 6. Gaud [GAU] | 13 | 10 | 4 | 15 | Eastern India | Indo-European | Middle Caste | Agriculture |
| 7. Gond [GND] | 51 | 10 | | | Central India - Primarily Madhya Pradesh | Dravidian - Gondi dialect | Tribe | Agriculture, food gathering and humting |
| 8. Halba [HAL] | 47 | 20 | 20 | 48 | Central India - Primarily Madhya Pradesh | Indo-European Primarily Marathi | Tribe | Agriculture, food gathering and humting |
| 9. Ho [HO] | 54 | 10 | 20 | 54 | Eastern India – Primarily Bihar | Austro-Asiatic | Tribe | Agriculture, food gathering and humting |
| 10. Irula [ILA] | 30 | 14 | 18 | 50 | Southern India - Primarily Tamilnadu, including Nilgiri Hills | Dravidian | Tribe | Shifting cultivation |
| 11. Iyengar [IYN] | 30 | 10 | 20 | 51 | Southern India - Tamilnadu | Dravidian | Upper Caste | Traditionally priests, now various occupations |
| 12. Iyer [IYR] | 30 | 10 | 20 | 50 | Southern India - Tamilnadu | Dravidian | Upper Caste | Traditionally priests, now various occupations |
| 13. Jamatiya [JAM] | 55 | 10 | 16 | 55 | North-eastern India - Primarily Tripura | Tibeto-Burman | Tribe | Agriculture |

... CONTINUED

TABLE 2.1 ... CONTINUED

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 14. Jat Sikh [JSK] | 48 | 15 | | | Northern India - Punjab | Indo-European | Middle Caste | Various occupations. including agriculture. |
| 15. Kamar [KMR] | 54 | 10 | 19 | 57 | Central India - Primarily Madhya Pradesh | Dravidian | Tribe | Agriculture, food gathering and humting |
| 16. Khatris [KHT] | 48 | 15 | | | Northern India - Punjab | Indo-European | Middle Caste | Various occupations. including agriculture. |
| 17. Konkan Brahmins [KBR] | 31 | 10 | | | Western India - Maharashtra | Indo-European | Upper Caste | Traditionally priests, now various occupations . |
| 18. Kota [KOT] | 30 | 25 | 15 | 45 | Southern India - Nilgiri Hills | Dravidian | Tribe | Artisans, Musici-ans and Agricul-turists |
| 19. Kurumba [KUR] | 30 | 10 | 18 | 54 | Southern India - Nilgiri Hills | Dravidian | Tribe | Hunting and food gathering |
| 20. Lodha [LOD] | 32 | 14 | 17 | 32 | Eastern India - West Bengal | Austro-Asiatic | Tribe | Hunting, food gathering and agricultural labor |
| 21. Mahishya [MAH] | 33 | 10 | 9 | 34 | Eastern India | Indo-European | Middle Caste | Agriculture |
| 22. Manipuri (Meitei) [MNP] | 11 | 9 | | | North-eastern India - Manipur | Tibeto-Burman | Upper caste | Agriculture and various occupations. |
| 23. Maratha [MRT] | 41 | 10 | | | Western India - Maharashtra | Indo-European | Middle Caste | Various occupations, including agriculture |
| 24. Mizo [MZO] | 29 | 14 | 20 | 29 | North-eastern India - Mizoram | Tibeto-Burman | Tribe | Agriculture, basket making |
| 25. Mog [MOG] | 25 | 10 | 6 | 25 | North-eastern India - Primarily Tripura | Tibeto-Burman | Tribe | Agriculture |
| 26. Munda [MUN] | 7 | 6 | | 49 | Eastern India | Austro-Asiatic | Tribe | Hunting, food gathering and agriculture |
| 27. Muria [MUR] | 30 | 12 | 8 | 28 | Central India - Primarily Madhya Pradesh | Dravidian - Gondi dialect | Tribe | Agriculture, food gathering and hunting |
| 28. Muslim [MUS] | 28 | 10 | 19 | | Throughout India | Indo-European | Islamic Religious Group | Various occupations, including agriculture |
| 29. Naba-Baudh [NBH] | 40 | 10 | | | Western India - Maharashtra | Indo-European | Lower caste (recently adopted Buddhism) | Various occupations, including agriculture |

... CONTINUED

TABLE 2.1 ... CONTINUED

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 30. Pallan [PLN] | 30 | 10 | 15 | 50 | Southern India - Tamilnadu | Dravidian | Lower Caste | Agriculture |
| 31. Punjab Brahmhins [PBR] | 48 | 12 | | | Northern India - Punjab | Indo-European | Upper Caste | Traditionally priests, now various occupations |
| 32. Rajput [RAJ] | 51 | 10 | 35 | 52 | Northern and Western India | Indo-European | Middle Caste | Various occupations, including agriculture |
| 33. Riang [RIA] | 51 | 12 | 17 | 50 | North-eastern India - Primarily Tripura | Tibeto-Burman | Tribe | Agriculture |
| 34. Santal [SAN] | 20 | 14 | 15 | 24 | Eastern India | Austro-Asiatic | Tribe | Agriculture, hunting and food gathering |
| 35. Saryupari Brahmins [SBR] | 26 | 19 | | | Central India – Madhya Pradesh | Indo-European | Upper Caste | Traditionally priests, now various occupations |
| 36. Scheduled caste – Punjab [SCH] | 48 | 15 | | | Northern India - Punjab | Indo-European | Lower Caste | Various occupations, including agriculture |
| 37. Tanti [TAN] | 16 | 10 | 6 | 16 | Eastern India | Indo-European | Lower Caste | Weaving and agricultural labor |
| 38. Tipperah (Tripuri) [TRI] | 51 | 20 | 17 | 50 | North-eastern India - Primarily Tripura | Tibeto-Burman | Tribe | Agriculture |
| 39. Toda [TOD] | 50 | 10 | 8 | 50 | Southern India - Primarily Tamilnadu, including Nilgiri Hills | Dravidian | Tribe | Hunting and food gathering |
| 40. Toto [TTO] | 30 | 20 | 12 | 30 | West Bengal- particularly Jalpaiguri district | Tibeto-Burman | Tribe | Agriculture |
| 41. Uttar Pradesh Brahmins [UBR] | 27 | 10 | 17 | 27 | Northern India - Uttar Pradesh | Indo-European | Upper Caste | Traditionally priests, now various occupations |
| 42. Vanniyar [VAN] | 30 | 10 | 14 | 50 | Southern India - Tamilnadu | Dravidian | Middle Caste | Traders, Agriculture |
| 43. Vellala [VLR] | 43 | 10 | 16 | 43 | Southern India - Tamilnadu | Dravidian | Middle Caste | Agriculture |
| 44. West Bengal Brahmins [WBR] | 22 | 10 | 13 | 23 | Eastern India - West Bengal | Indo-European | Upper Caste | Traditionally priests, now various occupations |

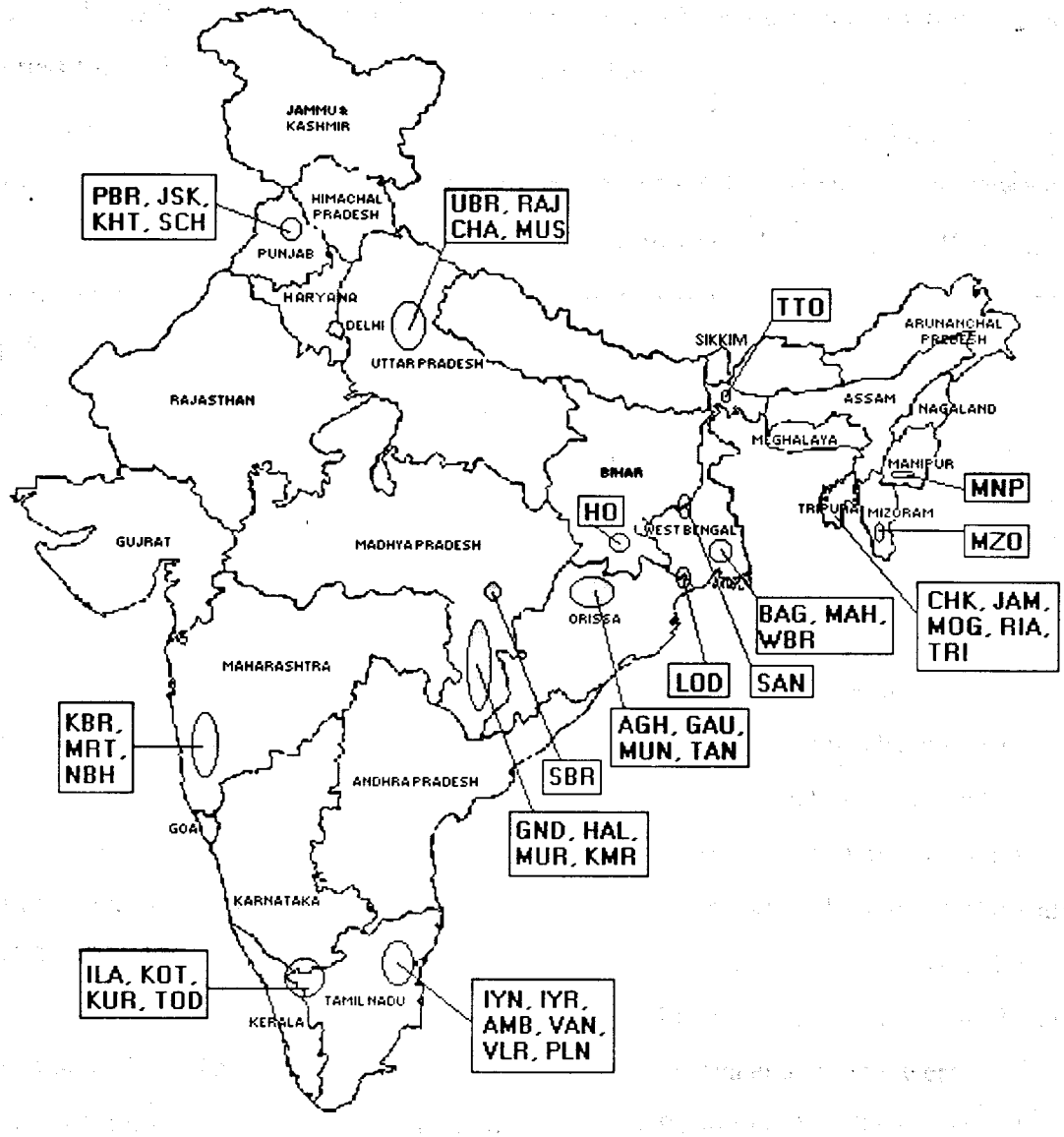



**Figure 2.1**
Map showing the geographical locations of sampling of the 44 ethnic populations of India included in the present study

from Y-chromosomal or autosomal DNA analyses. Since many of our statistical analyses necessitated pooling of populations, this limitation did not turn out to be serious. Sample sizes are given in Table 2.1. Because of failure of experiments, there are slight variations in sample sizes across loci, which are indicated in appropriate tables.

*Loci and Protocols*

Each DNA sample was screened for 10 mtDNA restriction site polymorphisms (RSPs) and 1 Insertion/Deletion polymorphism (IDP). The RSPs screened were HaeIII np 663, HpaI np 3592, AluI np 5176, AluI np 7025, DdeI np 10394, AluI np 10397, HinfI np 12308, HincII np 13259, AluI np 13262, HaeIII np 16517; and the IDP screened was the COII/tRNA$^{Lys}$ intergenic 9-bp deletion. These sites were chosen such that individuals could be classified into those haplogroups which, from past studies, are considered to be the most relevant for Indian populations. mtDNA RSP analyses were performed using standard primers and protocols (Torroni et al. 1993, 1996). Sequencing of the Hyperveriable Segment-1 (HVS1) of mitochondrial DNA (mtDNA) was carried out by cycle sequencing method in ABI-3100 automated DNA sequencer and the ABI prism dideoxyterminator system. The HVS1 region (np 16024 – np 16380) was amplified using standard primers in both directions (Vigilant et al 1991).

DNA samples were typed for 18 Y-chromosomal markers; 12 of which were binary polymorphic markers while 6 were short tandem repeat markers. The 12 binary markers are YAP, 92r7, SRY 4064, sY81, SRY+465, TAT, M9, M13, M17, M20, SRY10831 and p12f2. Primer sequences and amplification protocols for the first 11 DNA markers are described by Thomas et al. (1999). The primer sequences for p12f2 (Casanova et al.1985), were 12f2D and 12f2G and the amplification protocol was as described by Rosser et al. (2000) to amplify a 88-bp product. As an internal control, a product of size 148-bp encompassing the M172 polymorphism was co-amplified using the primers M172-F 5' - TCCCCCAAACCCATTTTGATGCAT - 3' and M172-R 5' - GGATCCATCTTCACTCAATGTTG - 3'. PCR amplification was carried out in 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 2.5 mM $MgCl_2$ , 0.2 mM of each dNTP, 0.2 $\mu$ M of each p12f2-primer, 0.3 $\mu$ M of each M172-primer, and 0.2 U of AmpliTaq Gold (Perkin Elmer, USA). Cycling conditions were as follows: initial denaturation for 5 min; 30 cycles of denaturation at 94°C for 30s, annealing at 58°C for 45s and extension at 72°C for 45s. The final cycle ended with an additional extension of 10 min at 72°C. The six short tandem repeat markers

were DYS19, DYS388, DYS390, DYS391, DYS392, DYS393; all of which were amplified using markers and protocols as described by Thomas et al. (1999). Restriction digested products and the amplified DNA were electrophoresed on an ABI-3100 automated DNA sequencer and genotyped by Genescan version 3.1 and Genotyper version 2.1. In this study we have used the haplogroup definitions as given in Rosser et al. (2000).

Each DNA sample was analysed for polymorphisms at 25 autosomal loci; of which 8 were insertion/deletion polymorphisms (IDPs) and remaining 17 were RFLPs. The names and GDB accession numbers or ALFRED UID of the RSP loci are: ESR1 (GDB:185229); NAT (GDB:187676); CYP1A (GDB: 9956062)-MspI; PSCR (GDB:182305); T2 (GDB:196856); LPL (GDB:285016); ALB (GDB:178648); ALAD-*MspI* (GDB:155925); ALAD-*RsaI* (GDB:155924); HB $\psi\beta$ - *Hinc II* (GDB: 56084); HB 3'$\psi\beta$ - *HincII*; HB 5' $\beta$ - *HinfI*; HoxB4-*MspI* (UID: SI0001670); DRD2 (UID: SI000191L) - TaqIB, TaqID, Taq1A; ADH2-*RsaI* (UID: SI000002C). The names of the IDPs are given in Table 2.13. Primers and protocols used for screening of the IDPs were as given in Majumder et al. (1999a) and Tishkoff et al. (1996), and those for RSPs were as given in Jorde et al. (1995), Majumder et al. (1999b) and K. Kidd (personal communication).

The wet-laboratory experiments were not performed by me, but by others under the supervision of Professor Partha P. Majumder. I was responsible for data-cleanup, data-management and statistical analyses. However, I have provided details of the laboratory protocols for completeness.

## *Statistical Methods*

Allele frequencies at each locus were estimated for each population by the maximum likelihood method. Chi-squared tests of significance between the observed genotype frequencies and those expected under Hardy-Weinberg equilibrium were performed. Maximum likelihood estimates of haplotype frequencies at linked autosomal loci were obtained via the EM algorithm using the HAPLOFREQ package (Majumdar and Majumder 2000). Observed heterozygosities were estimated. Alignment of DNA sequences was done using CLUSTAL-w. The extent of genetic differentiation, $F_{ST}$, was estimated (Nei 1973; Hudson et al. 1992).

Tests of significance in non-sparse cross-classified tables were carried out by standard contingency chi-squared tests. For sparse tables, conventional statistical tests could not be used.

We have, therefore, used the following bootstrap test procedure. In a $k \times l$ frequency table, Let $n_{ij}$ denote the frequency in the (*i*,j)-th cell (*i*=1,2,...,*k*; *j*=1,2,...,*l*). Let $n_{.j} = \sum_{i=1}^{k} n_{ij}$ , $n_{i.} = \sum_{j=1}^{l} n_{ij}$ and $n = \sum_{i=1}^{k} \sum_{k=1}^{l} n_{ij}$ . We computed, for a pair of rows (or columns, depending on the hypothesis to be tested), the Bhattacharya's distance (Bhattacharya, 1946) , $d_{ij}$ , based on the observed proportions of the various haplotypes in these two populations. Then, we computed a statistic D $= \sum_{s<t} \frac{n_s + n_t}{n} \times d_{st}$ . To test the significance of D, we generated bootstrapped samples. For the s-th row (or, column), this was done by drawing a sample of size $n_s$ from a k-nomial distribution with cell probabilities $p_i = n_{i.} / n_{..}$ . When bootstrapped samples were so generated for all the rows, we had a bootstrapped $k \times l$ table. A large number (10,000) of such bootstrapped tables were generated. For each table, the D-statistic was calculated. The D-values were then sorted in ascending order and cut-off point corresponding to upper 5% tail were calculated. If the D-value corresponding to the observed table was above this cut-off point, then the null hypothesis was rejected at the 5% level of significance.

Stepwise logistic regression and linear discriminant analyses was carried out using SPSS. A non-parametric test, Kruskal-Wallis test, was performed to test equality of frequency distributions at Y-STRP loci, using SPSS. AMOVA analysis of mtDNA haplotypes was performed using ARLEQUIN (Schneider et al. 2000). DNA sequences were aligned using ClustalW. The Cambridge sequence was used as reference during alignment. Descriptive statistics, nucleotide diversities and mismatch statistics were calculated using the DnaSP (version 3) package (Rozas and Rozas 1999). Expansion times were estimated using the methodology proposed by Slatkin and Hudson (1991) assuming a mutation rate of 20.5% per site per million years which is appropriate for the HVS1 region (Bonatto and Salzano 1997). 95% confidence interval of an estimated expansion time was calculated using 2 × s.d. of the sampling variance of nucleotide diversity. Calculation of Fu's (1997) $F_S$ statistic and its test of significance using coalescent simulation were performed using ARLEQUIN (Schneider et al. 2000). For phylogenetic analyses using mtDNA sequence data, DNA distances were calculated using the maximum likelihood method assuming a 30:1 transition:transversion ratio, which has been

suggested as appropriate for the HVS1 region (Lundstrom et al. 1992). For phylogenetic analysis using frequencies of autosomal markers, mtDNA haplotypes and Y-chromosomal data, genetic distances were calculated using Nei's (1987) $D_A$ distance. The neighbor-joining method (Saitou and Nei 1987) was used for phylogenetic reconstruction. All phylogenetic calculations were performed using the DISPAN package or the DNADIST (Jukes-Cantor) and NEIGHBOR modules of the PHYLIP-ver. 3.5c package,

The age (A) of a Y-haplogroup were estimated as: $A = g \times s^2/\mu$, where g = generation time (assumed to be 30 years); $s^2$ = variance of STR repeat number among haplotypes belonging to the haplogroup, averaged over all 6 STR loci; and $\mu$ = mutation rate per generation at an STR locus (taken to be 0.18%, as previously estimated (Quintana-Murci et al. 2001) for the 6 STR loci under consideration). The 95% confidence interval of estimated A was calculated from the previously estimated (Quintana-Murci et al. 2001) 95% CI of $\mu$ = (0.31% - 0.098%).

An analysis of population structure using a Markov Chain Monte Carlo method as developed by Pritchard et al. (2000) was carried out using the program STRUCTURE.

### *Electronic Database Information*

ALFRED, http://alfred.med.yale.edu

ARLEQUIN, lgb.unige.ch/arlequin/

CLUSTAL-W, http://www2.ebi.ac.uk/clustalw/

DISPAN, http://oat.bio.indiana.edu:7580/

DnaSP, www.ub.es/dnasp/

Ethnologue, http://www.ethnologue.com/

*HVS1 database*, www.eva.mpg.de

NETWORK 3.0, http://www.fluxus-engineering.com/sharenet.htm

PHYLIP,http://evolution.genetics.washington.edu/phylip.html

STRUCTURE, http://pritch.bsd.uchicago.edu/software.html

## Results

### *Mitochondrial DNA Polymorphisms*

RSP haplotype frequencies: All populations are monomorphic at the *Hpa* I np 3592 locus. The *Dde* I np 10394, *Alu* I np 10397 and *Hae* III np 16517 loci are polymorphic in all populations.

Several populations are monomorphic at the *Hae* III np 663, *Alu* I np 5176, *Alu* I np 7025, *Hin*fI np 12308, *Hin*cII np 13259 and *Alu* I np 13262 loci. For the 9-bp COII / $tRNA^{lys}$ intergenic length mutation, no variation was observed in 40 out of the 44 populations. This length mutation was observed only among Riang (2 deletions among 51 individuals), Halba (1 deletion among 47 individuals), Gond (2 deletions among 51 individuals) and Nava Baudh (one insertion among 40 individuals). Although in samples drawn from a different geographical location, several individuals of the Irula tribe were found to poseess the 9-bp deletion (Watkins et al. 1999), we have not detected any in this population. The 9-bp deletion has also been reported to have arisen independently in India (Watkins et al. 1999). There is considerable variability (0.0054-0.4101) in average heterozygosities across the polymorphic loci. The *Alu* I np 10397 locus exhibited the highest heterozygosity of 0.4101, and the 9-base pair deletion locus showed the minimum heterozygosity (0.0054).

Table 2.2 presents the observed haplotype frequencies in the populations. Thirty two distinct haplotypes were observed among the 1490 individuals examined from the 44 populations. However, in none of the populations were all the 32 haplotypes observed. The maximum number of haplotypes (13) was observed among Rajput and Tipperah while the Kota and Toda harbored only two haplotypes each. The frequency distributions of the haplotypes among the populations are significantly different at the 5% level. However, one haplotype, 00111101010 accounted for 46.4% of all mtDNA molecules. This modal haplotype in the pooled data set is also the modal haplotype in 34 of the 44 study populations. It can, therefore be inferred that, this is the most ancient haplotype in Indian populations. The 10 populations, in which this haplotype is not the most frequent, primarily comprised ethnic groups of either the northern region (Uttar Pradesh Brahmins, Punjab Brahmins, Rajput, Muslim) or the northeastern region (Chakma, Jamatiya, Mog, Toto). This is consistent with known historical immigrations into these (north and northeastern) regions (Thapar 1966). In spite of the extensive haplotype sharing among all socio-culturally, geographically, and linguistically distinct ethnic populations, some haplotypes (00000001010, 00010000100, 00010001010, 00011001000, 00011001010, 00110001011) were found exclusively among northeastern Tibeto-Burman speaking tribal populations. This is consistent with the hypothesis of recent entry of these people and/or admixture with populations, possibly of southern China and adjoining regions, predominantly resident outside of India (Guha 1935; Su et al. 2000). We also analyzed these data after

Table 2.2

Estimated Percentage Frequencies of Haplotypes Based on 11 mtDNA Loci*

| SL. NO. | HAPLOTYPE | AGH | AMB | BAG | CHA | CHK | GAU | GND | HAL | HO | ILA | IYN | IYR | JAM | JSK | KBR | KHT | KMR | KOT | KUR | LOD | MAH | MNP | MOG | MRT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 00000001010 | | | | | | | | | | | | | | | | | | | | | | | 4.00 | |
| 2 | 00010000100 | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | 00010001000 | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | 00010001010 | | | | | | | | | | | | | | | | | | | | | | | 4.00 | |
| 5 | 00011001000 | | | | | 10.00 | | | | | | | | | | | | | | | | | | | |
| 6 | 00011001010 | | | | | | | | | | | | | | | | | | | | | | | 4.00 | |
| 7 | 00011101000 | | | | 4.00 | | | | | | | | | 5.45 | | | | | | | | | | 8.00 | |
| 8 | 00011101010 | | | | 4.00 | | | | | | | | | | | | | | | | | | | 48.00 | |
| 9 | 00100001000 | | | | | | | | | | | | 3.33 | | 2.08 | | 6.25 | | | | | | | | |
| 10 | 00100001010 | | | | | | | 1.96 | 2.13 | | | | | | 2.08 | | 4.17 | | | | | | | | |
| 11 | 00100011000 | | | | | | | | | | | | | | 2.08 | | | | | | | | | | |
| 12 | 00110000110 | | | | | | | | | | | | | | 4.17 | | 2.08 | | | | | | | | 2.44 |
| 13 | 00110001000 | 4.17 | 3.33 | 3.23 | | | 15.38 | | 2.13 | 3.70 | | 3.33 | 6.67 | | 8.33 | 3.23 | 8.33 | 9.26 | | | | | 9.09 | | 9.76 |
| 14 | 00110001010 | | | 19.35 | 4.00 | 50.00 | 7.69 | 15.69 | 17.02 | 14.81 | 23.33 | 33.33 | 13.33 | 36.36 | 18.75 | 6.45 | 12.50 | 7.41 | 3.33 | 13.33 | | 24.24 | 36.36 | 20.00 | 14.63 |
| 15 | 00110001011 | | | | | | | | | | | | | | | | | | | | | | | | |
| 16 | 00110011000 | 12.50 | 6.67 | 6.45 | 8.00 | | 7.69 | 3.92 | 4.26 | 1.85 | 20.00 | 10.00 | 10.00 | | 6.25 | 3.23 | 10.42 | 7.41 | | | | 12.12 | 9.09 | | 2.44 |
| 17 | 00110011003 | | | | | | | | | | | | | | | | | | | | | | | | |
| 18 | 00110011010 | 8.33 | | 6.45 | 24.00 | | 7.69 | 9.80 | 2.13 | 5.56 | 3.33 | | 13.33 | 1.82 | 10.42 | 3.23 | 10.42 | | | 6.67 | 18.75 | 6.06 | | | 4.88 |
| 19 | 00111001000 | | | | | | | | | | | 3.33 | | | | | | | | | | | | | |
| 20 | 00111001010 | | | 3.23 | | 10.00 | | | 2.13 | | | | | | 4.17 | | | | | | | 3.03 | | | 2.44 |
| 21 | 00111011010 | | | | | | 15.38 | | | | | | | | 2.08 | | | | | | | | | | |
| 22 | 00111100110 | | | | | | | | | | 3.33 | | 3.33 | | | 9.68 | | | | | | | 9.09 | | |
| 23 | 00111101000 | 4.17 | 33.33 | 12.90 | | 10.00 | 7.69 | 3.92 | 6.38 | 20.37 | 3.33 | | | 18.18 | 4.17 | | 4.17 | 16.67 | | | 34.38 | 3.03 | | 4.00 | 9.76 |
| 24 | 00111101001 | | | | | | | 1.96 | | | | | | | | | | | | | | | | | |
| 25 | 00111101010 | 70.83 | 56.67 | 48.39 | 56.00 | 20.00 | 30.77 | 60.78 | 61.70 | 53.70 | 46.67 | 50.00 | 46.67 | 34.55 | 35.42 | 70.97 | 41.67 | 59.26 | 96.67 | 76.67 | 46.88 | 42.42 | 36.36 | 8.00 | 53.66 |
| 26 | 00111101011 | | | | | | | 1.96 | 2.13 | | | | | | | | | | | | | | | | |
| 27 | 00111111010 | | | | | | 7.69 | | | | | | | | | 3.23 | | | | | | | | | |
| 28 | 10010001000 | | | | | | | | | | | | | | | | | | | | | | | | |
| 29 | 10110001000 | | | | | | | | | | | | 3.33 | | | | | | | | | 6.06 | | | |
| 30 | 10110001010 | | | | | | | | | | | | | 3.64 | | | | | | 3.33 | | | | | |
| 31 | 10110011000 | | | | | | | | | | | | | | | | | | | | | 3.03 | | | |
| 32 | 10111101000 | | | | | | | | | | | | | | | | | | | | | | | | |

... continued

## Table 2.2 ... continued

| SL. NO. | HAPLOTYPE | MUN | MUR | MUS | MZO | NBH | PBR | PLN | RAJ | RIA | SAN | SBR | SCH | TAN | TOD | TRI | TTO | UBR | VAN | VLR | WBR | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 00000001010 | | | | | | | | | | | | | | | | | | | | | 0.07 |
| 2 | 00010000100 | | | | | | | | | 1.96 | | | | | | 3.92 | | | | | | 0.20 |
| 3 | 00010001000 | | | | | | | | | | | | | | | | | 3.70 | | | | 0.07 |
| 4 | 00010001010 | | | | | | | | | | | | | | | | 3.33 | | | | | 0.13 |
| 5 | 00011001000 | | | | | | | | | | | | | | | | | | | | | 0.07 |
| 6 | 00011001010 | | | | | | | | | | | | | | | | | | | | | 0.07 |
| *7 | 00011101000 | | | 7.14 | 3.45 | | | | | | | | | | | 3.92 | 53.33 | | | | | 1.81 |
| 8 | 00011101010 | | | | | | | | 5.88 | 7.84 | | | | | | 3.92 | 26.67 | 3.70 | | | | 2.08 |
| 9 | 00100001000 | | | | | | 2.08 | | | | | | 6.25 | | | | | 3.70 | | | | 0.67 |
| 10 | 00100001010 | | | | | | 2.08 | | | | | | | | | | | 3.70 | | | | 0.47 |
| 11 | 00100011000 | | | | | | | | | | | | | | | | | | | | | 0.07 |
| 12 | 00110000110 | | | | | | 2.08 | | 1.96 | | | | 2.08 | | | 1.96 | | 3.70 | | | | 0.60 |
| 13 | 00110001000 | | | 3.57 | | 2.50 | 10.42 | | 9.80 | 3.92 | 5.00 | 3.85 | 4.17 | | | 1.96 | | | 16.67 | | 9.09 | 3.76 |
| 14 | 00110001010 | 28.57 | 13.33 | 10.71 | 17.24 | 15.00 | 29.17 | 20.00 | 25.49 | 11.76 | 15.00 | 11.54 | 4.17 | 6.25 | 72.00 | 21.57 | | 37.04 | 20.00 | 41.86 | 13.64 | 18.72 |
| 15 | 00110001011 | | | | | | | | | 3.92 | | | | | | | | | | | | 0.13 |
| 16 | 00110011000 | | 3.33 | 3.57 | 3.45 | 2.50 | 4.17 | 3.33 | 3.92 | | 10.00 | 7.69 | 10.42 | 6.25 | | 1.96 | | 7.41 | | 2.33 | | 4.63 |
| 17 | 00110011003 | | | | | 2.50 | | | | | | | | | | | | | | | | 0.07 |
| 18 | 00110011010 | | 3.33 | 28.57 | | 12.50 | 16.67 | 3.33 | 11.76 | | | 11.54 | 18.75 | 6.25 | | 3.92 | | 18.52 | 10.00 | | 4.55 | 6.85 |
| 19 | 00111001000 | | | 3.57 | | | 2.08 | | 1.96 | | | 3.85 | | | | | | | | | | 0.34 |
| 20 | 00111001010 | | | | 6.90 | | 2.08 | | 5.88 | | | 3.85 | | | | | | | | | 13.64 | 1.14 |
| 21 | 00111011010 | | | | | | 4.17 | | 1.96 | | | | 4.17 | | | | | | | | | 0.54 |
| 22 | 00111100110 | | | | 3.45 | | | | 3.92 | 1.96 | | | | | | 1.96 | | | | | | 0.74 |
| 23 | 00111101000 | 28.57 | 3.33 | 3.57 | 10.34 | 7.50 | 2.08 | 10.00 | 1.96 | 7.84 | 10.00 | 7.69 | 2.08 | 6.25 | | 9.80 | 3.33 | | 13.33 | 16.28 | 4.55 | 7.85 |
| 24 | 00111101001 | | | | | | | | | | | | | | | | | | | | | 0.07 |
| 25 | 00111101010 | 42.86 | 76.67 | 25.00 | 48.28 | 57.50 | 22.92 | 56.67 | 19.61 | 47.06 | 60.00 | 50.00 | 47.92 | 75.00 | 28.00 | 35.29 | 10.00 | 14.81 | 40.00 | 39.53 | 54.55 | 46.44 |
| 26 | 00111101011 | | | | | | | | | | | | | | | | | | | | | 0.13 |
| 27 | 00111111010 | | | | | | | 6.67 | | | | | | | | 1.96 | | | | | | 0.34 |
| 28 | 10010001000 | | | | | | | | | 3.92 | | | | | | | | | | | | 0.13 |
| 29 | 10110001000 | | | 14.29 | 6.90 | | | | 5.88 | 9.80 | | | | | | 7.84 | | 3.70 | | | | 1.48 |
| 30 | 10110001010 | | | | | | | | | | | | | | | | | | | | | 0.20 |
| 31 | 10110011000 | | | | | | | | | | | | | | | | | | | | | 0.07 |
| 32 | 10111101000 | | | | | | | | | | | | | | | | 3.33 | | | | | 0.07 |

* Order of loci: HaeIII np 663, HpaI np 3592, AluI np 5176, AluI np 7025, DdeI np 10394, AluI np 10397, HinfI np 12308, HinclI np 13259, AluI np 13262, HaeIII np 16517, 9-bp deletion. (1= presence of restriction site/deletion, 0= absence of restriction site/deletion)

grouping populations by language, geographical region or social rank. The Austro-Asiatic speaking tribal populations harbored the minimum number (6) of haplotypes, while the Dravidian, Tibeto-Burman and Indo-European speaking populations harbored, respectively 15, 22, and 22 distinct haplotypes. The frequency of the modal haplotype (00111101010) is significantly ($p<0.0001$) lower among the Tibeto-Burman (32%) speaking groups, while those among the Austro-Asiatic, Dravidian or Indo-European speaking groups are not significantly ($p>0.05$) different. This modal haplotype is also modal across tribal and caste groups of all ranks, though not among the Muslim. However, the difference in frequencies of the modal haplotype between tribals are castes is statistically significant ($p<.00005$). This indicates that the frequency of the major ancestral female lineage have been altered by subsequent admixture and/or drift. The modal haplotype among the Muslim is, interestingly, on the U haplogroup background (described in detail later), which is also quite frequent among the caste, but not among tribal, groups. The modal haplotype is the same across all geographical regions, although the frequencies among the regions are variable and statistically significantly different [62% (central region) – 33% (northern and north-eastern regions)]. The second most frequent haplotype is 00110001010 in all the linguistic groups, except the Austro-Asiatic. This haplotype is on a non-M haplogroup background (described in detail later), while among the Austro-Asiatic the second most frequent haplotype is different (00111101000) and on a M background. However, even though the second most frequent haplotype among Dravidian, Tibeto-Burman and Indo-European is the same, the frequencies of this haplotype are significantly ($p<0.05$) different among these linguistic groups. The haplotype 00110011010, on a non-M background, occurs with a high frequency (16.1%) in the northern region, but is significantly lower (1.1% - 7.1%) in the other regions. Another haplotype, 00111101000, occurs with a high frequency (13.9%) in the eastern region, but occurs at lower frequencies in the other regions (2.5% - 9.5%). Thus, the conclusion from the above findings is that there is strong evidence of a fundamental unity of female lineages across ethnic populations of India, irrespective of their geographical location of habitat, socio-cultural or linguistic affiliation. The minor deviations that are found to occur in certain populations/locations are easily explained by documented historical immigrations (such as of Indo-European speakers into northern India).

Population-specific haplotype diversities are presented in Figure 2.2. The haplotype diversity in most populations, except the Kota, is quite high (64.30%-89.74%).

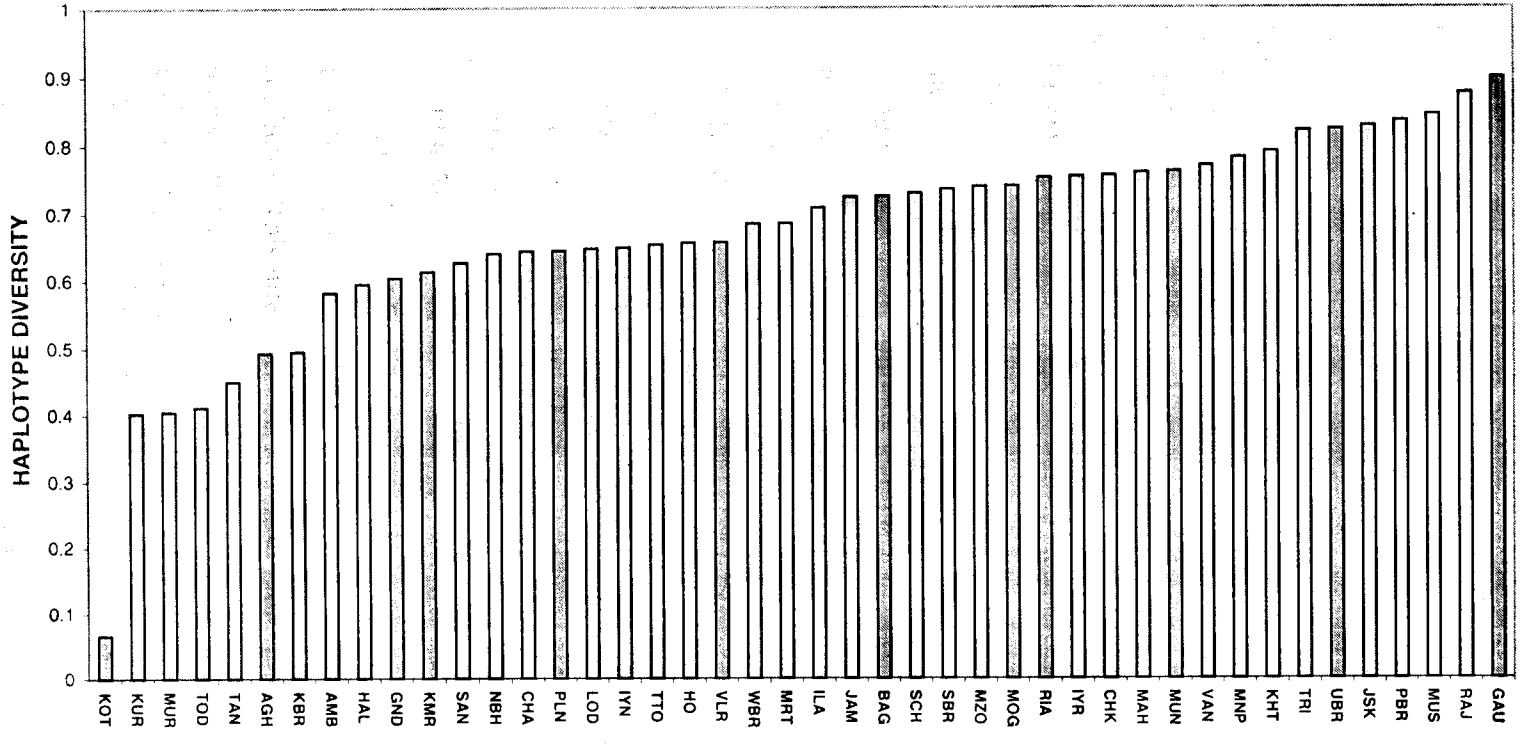



**Figure 2.2**
Diversities of mitochondrial RSP haplotypes among 44 ethnic populations of India

Distribution of haplogroups: The RSPs examined in the present study permitted the classification of individuals into the following haplogroups (HGs): A, B, C, D, H, L, M, U. Table 2.3 presents the frequencies of these haplogroups in the 44 study populations. HG-L was not found in any of the study populations. HGs A and C, which are predominant haplogroups in Siberia and among American Indians (Wallace 1995), occur at intermediate frequencies (1.96-%-14.20%) in study populations. HG-B, which is present in very high frequencies in the Pacific Islands (Wallace 1995), is present only among the Riang with 4% frequency. HG- D is present mostly in the northeastern populations; the highest frequency was observed among the Mog (56%). This haplogroup has been reported to be frequent in populations of Tibet and Korea, among the Han Chinese and also among American Indians (Wallace 1995). Of particular interest are the frequencies of HGs M and U. HG-M has been proposed to be an ancient east-Asian marker (Ballinger et al 1992) and is virtually absent among African (except Ethiopian) and Caucasoid populations (Torroni et al 1994; Passarino et al 1996a; Passarino et al 1996b). The origin of HG-M has been somewhat controversial. Quintana-Murci et al (1999) have proposed that the origin of HG-M is in Africa, while others (Roychoudhury et al. 2001; Maca-Meyer et al. 2001) have proposed that it may have originated in India and later migrated to Africa. HG-U has been found in high frequencies among Caucasoid populations, making it suitable for identifying Caucasoid admixture in Indian populations. The frequency of HG-M in Indian populations is very high (overall 59.9%: range 18.5% [Brahmins of Uttar Pradesh] to 96.7% [Kota]), confirming that it is an ancient marker in India. HG-M frequencies are significantly different ($p<.001$) among the geographical regions. It is lowest among north Indian populations (38.39%). From north India, HG-M frequency increases towards all other directions (Figure 2.3). It is possible that the frequency of HG-M was uniformly high throughout India, before the arrival of the Indo-European speakers from Central Asia about 3000-4000 ybp through the northwest corridor of India. Subsequently, as these Indo-European speakers penetrated into India, they pushed back the existing populations, thereby establishing a gradient of HG-M frequencies from the north towards the other regions of India. This inference, of course, assumes that the Indo-European speakers who entered India were not exclusively male; there must also have been a significant number of females with non HG-M mitochondrial haplotypes. HG-M frequencies are significantly different ($p<.001$) among the populations belonging to different socio-cultural groups. It is highest among the tribes followed by lower caste, middle caste, upper caste and

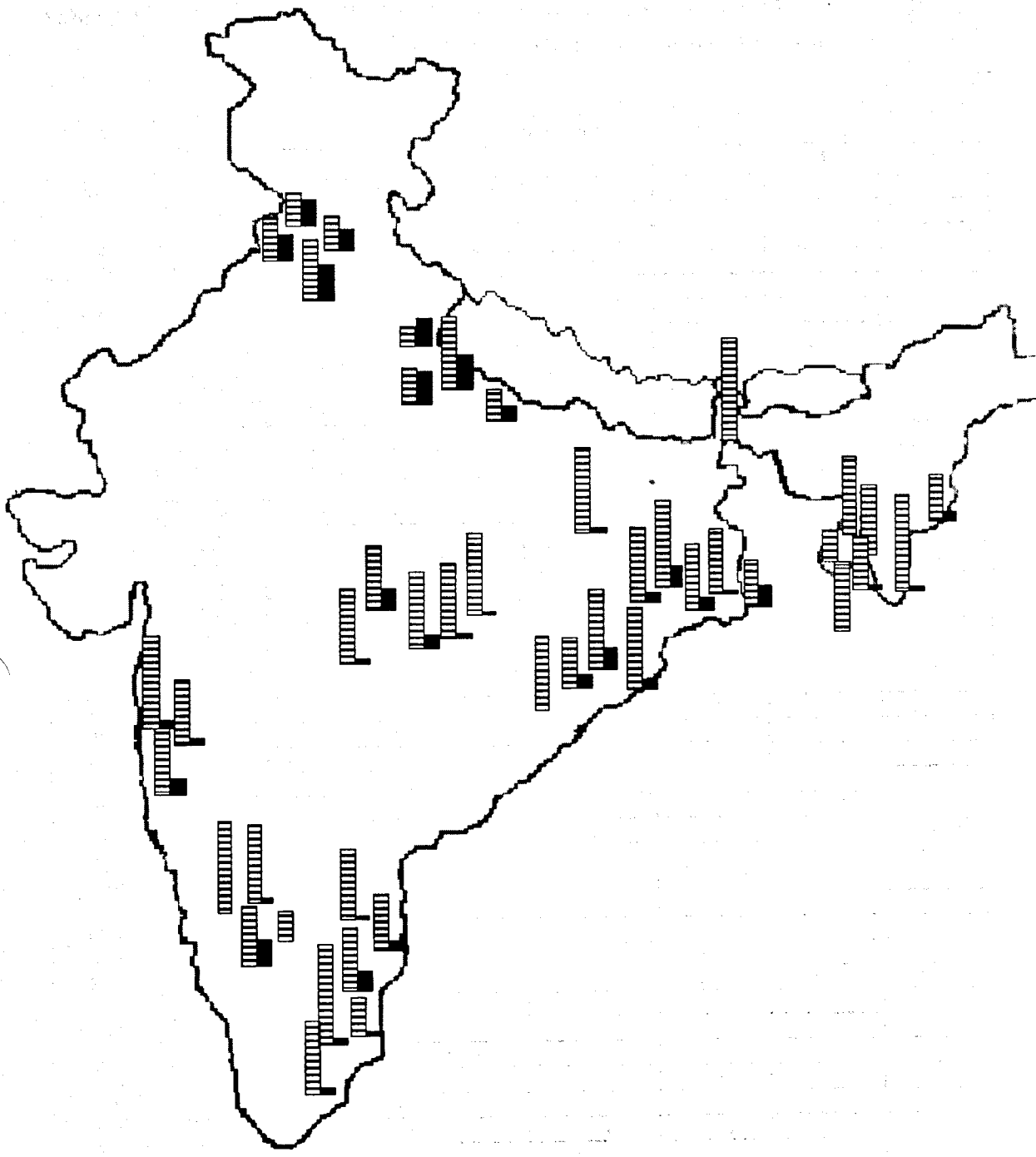


**Figure 2.3**
Frequencies (%) of mitochondrial haplogroups M (hatched ) and U (solid black) in 44 ethnic populations of India

**Table 2.3**
**Percentage Frequencies of Various Known Haplogroups in 44 Ethnic Populations of India and in the Total Sample**

| Population Code | Haplogroups* | | | | | | |
|---|---|---|---|---|---|---|---|
| | M | U | H | D | C | B | A |
| AGH | 75.00 | 20.83 | | | | | |
| AMB | 90.00 | 6.67 | | | | | |
| BAG | 61.29 | 12.90 | | | | | |
| CHA | 64.00 | 32.00 | | 8.00 | | | |
| CHK | 30.00 | | | | | | |
| GAU | 46.15 | 15.38 | | | | | |
| GND | 68.63 | 13.73 | 1.96 | | | | |
| HAL | 70.21 | 6.38 | 2.13 | | | | |
| HO | 74.07 | 7.41 | | | | | |
| ILA | 53.33 | 23.33 | | | 3.33 | | |
| IYN | 50.00 | 10.00 | | | | | |
| IYR | 50.00 | 23.33 | 3.33 | | 3.33 | | 3.33 |
| JAM | 58.18 | 1.82 | | 5.45 | | | 3.64 |
| JSK | 39.60 | 18.70 | 6.25 | | | | |
| KBR | 83.87 | 6.45 | | | 9.68 | | |
| KHT | 45.80 | 20.80 | 10.40 | | | | |
| KMR | 75.93 | 7.41 | | | | | |
| KOT | 96.67 | | | | | | |
| KUR | 76.67 | 6.67 | | | | | 3.33 |
| LOD | 81.25 | 18.75 | | | | | |
| MAH | 45.45 | 21.21 | | | | | 9.09 |
| MNP | 45.45 | 9.09 | | | 9.09 | | |
| MOG | 68.00 | | 4.00 | 56.00 | | | |
| MRT | 63.41 | 7.32 | | | | | |
| MUN | 71.43 | | | | | | |
| MUR | 80.00 | 6.67 | | | | | |
| MUS | 35.71 | 32.14 | | 7.14 | | | 14.29 |
| MZO | 65.52 | 3.45 | | 3.45 | 3.45 | | 6.90 |
| NBH | 65.00 | 17.50 | | | | | |
| PBR | 25.00 | 20.80 | 4.16 | | | | |
| PLN | 73.33 | 6.67 | | | | | |
| RAJ | 31.37 | 15.69 | | 5.88 | 3.92 | | 5.88 |
| RIA | 64.71 | | | 7.84 | 1.96 | 3.92 | 13.73 |
| SAN | 70.00 | 10.00 | | | | | |
| SBR | 57.69 | 19.23 | | | | | |
| SCH | 50.00 | 29.16 | 6.25 | | | | |
| TAN | 81.25 | 12.50 | | | | | |
| TOD | 28.00 | | | | | | |
| TRI | 56.86 | 5.88 | | 7.84 | 1.96 | | 7.84 |
| TTO | 96.67 | | | 46.67 | | | |
| UBR | 18.52 | 25.93 | 7.41 | 3.70 | | | 3.70 |
| VAN | 53.33 | 10.00 | | | | | |
| VLR | 55.81 | 2.33 | | | | | |
| WBR | 59.09 | 4.55 | | | | | |
| Total | 59.46 *(n=886)* | 11.67 *(n=174)* | 1.27 *(n=19)* | 3.22 *(n=48)* | 0.74 *(n=11)* | 0.13 *(n=2)* | 1.88 *(n=28)* |

* Haplogroups C and D are subsets of haplogroup M; therefore individuals belonging to haplogroups C and D are also counted as belonging to haplogroup M. Individuals belonging to haplogroup A, B, H and U are all non-M.

Muslims. The frequencies of HG-M in different categories are depicted in Figure 2.4(a). The frequencies of this haplogroup are also significantly different ($p<.001$) among the four linguistic groups. It is found that Austro-Asiatic speaking group harbors the highest HG-M frequency, which is significantly higher ($p<0.03$) than the Dravidian, Tibeto-Burman and Indo-European groups. Among Indo-Europeans, the HG-M frequency is significantly lower ($p< 0.001$) than Dravidians and Tibeto-Burmans.

There is an inverse relationship, albeit somewhat rough, between the frequencies of the M and U haplogroups (Figure 2.3). HG-U frequency is significantly higher ($p<.001$) among the north Indian populations than among populations of other geographical regions. Thus, the Indo-European speakers in India harbor the highest frequency of HG-U; this value is significantly higher ($p<.001$) than among Austro-Asiatic, Dravidian and Tibeto-Burman speakers (Figure 2.4(b)). Among the caste populations, although the lower caste groups have the highest HG-U frequency, the frequencies of this haplogroup are not significantly different among the caste groups of different ranks. Another relevant Caucasian specific haplogroup is HG-H. This haplogroup occurs at high frequencies among north Indian populations (Uttar Pradesh Brahmin, Punjab Brahmin, Khattri). These findings are consistent with our earlier inference of Indo-European speakers geographically displacing pre-existing populations.

In order to determine the relative effects of linguistic, socio-cultural and geographical differences among the populations on HG-M and U frequencies, we carried out a stepwise logistic regression analysis. We found that all of the three factors are significant in explaining the observed variation in the frequencies of both HGs M and U, although rankings of their relative effects differ for the two haplogroups. For HG-M, geographical locations of the study populations have the most significant effect, followed by linguistic grouping and socio-cultural status. On the other hand, for HG-U, linguistic grouping has the most significant effect followed by geographical zone and socio-cultural category. In view of the confounding of language and geographical groupings mentioned earlier, we think that these results reinforce our earlier interpretation that HG-M was generally ubiquitous in India and with the infusion of HG-U lineages by immigration of Indo-European speakers and the resultant pushback of the pre-existing populations in the southern and eastern directions, there is clear geographical and linguistic distinction in frequencies of these haplogroups.

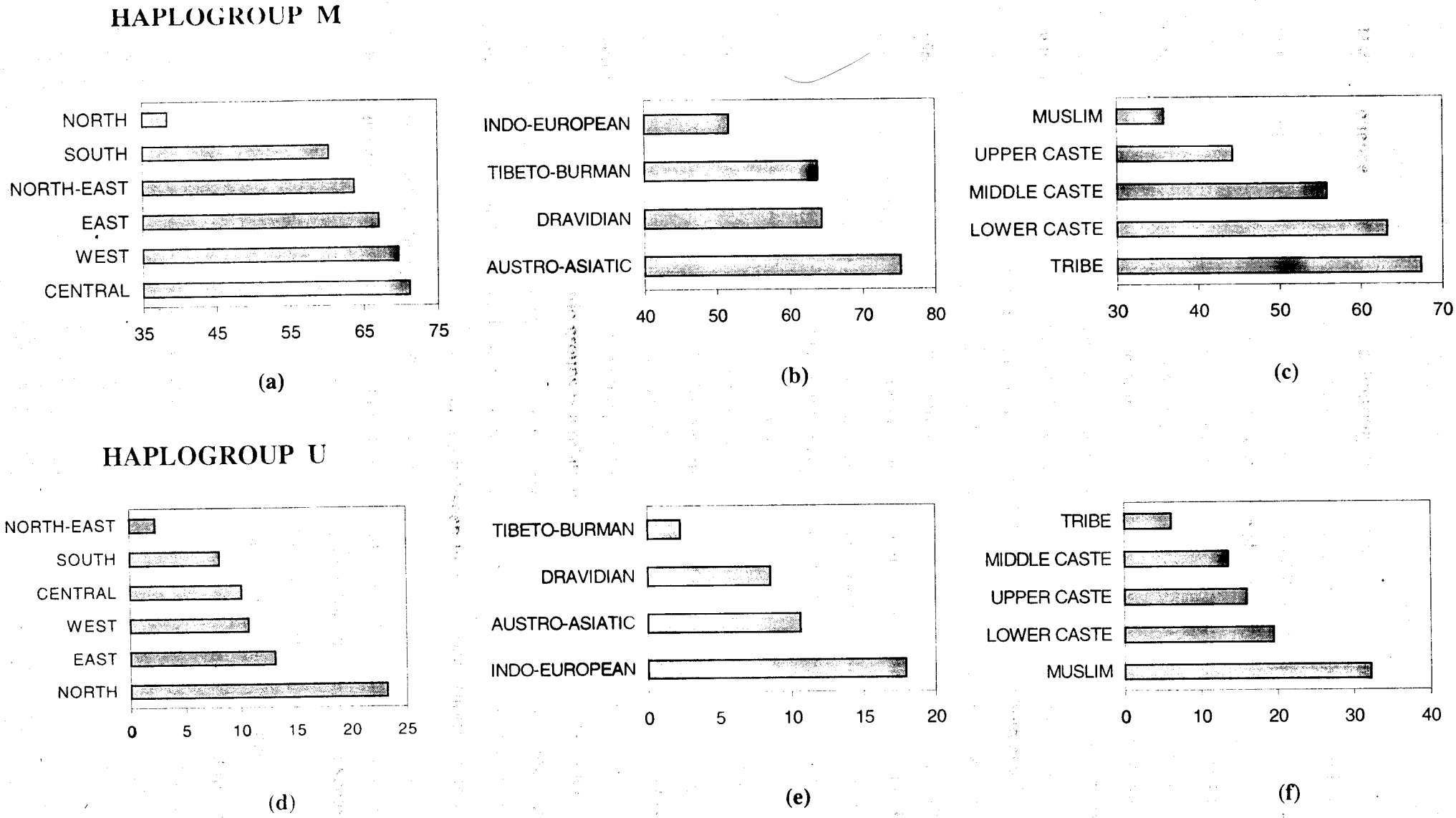



**Figure 2.4**
Frequencies (%) of mitochondrial haplogroups M and U in populations belonging to various geographical (*a* and *d*), linguistic (*b* and *e*) and socio-cultural (*c* and *f*) groups.

HVS1 sequence variation and subhaplogroup frequencies: The hypervariable segment 1 (HVS1) was sequenced in a total of 528 individuals, randomly selecting at least 10 individuals from each ethnic group. Upon alignment of these HVS1 sequences against the CRS (Cambridge Reference Sequence), four gaps were introduced in the CRS: one after np 16169, one after 18183 and two after 18189. These four gaps were eliminated from all the sequences prior to statistical analysis. The stretch of 357 nucleotides of HVS1 region, show deletions of nucleotides at positions 16166, 16182, 16183, 16189, 16190 in several individuals and there are a total 153 polymorphic sites. Table 2.4 provides the list of observed frequencies of various nucleotides at the 153 positions at which variant nucleotides (compared to the CRS) were noted in the pooled sample.

The number of distinct sequences among the 528 individuals for whom HVS1 sequencing was carried out is 323. Of the 323 distinct sequences, 91 (28.2%) sequences were shared by 296 (56.1%) of the 528 individuals. Thus, there is considerable sharing of HVS1 sequences, again reinforcing our earlier inference of a fundamental unity of female lineages in India. It was of interest to investigate the nature of sharing of sequences across the various groupings of populations (Figure 2.5). From this figure, it is clear that there is considerable sharing of sequences among tribals and castes. In the subset of sequences that are shared across socio-cultural groups, the average extent of sharing is 3.67 individuals per sequence. This average number is, interestingly, lower (2.89) for the subset of sequences that are shared by individuals only within a particular socio-cultural group. Thus, the extent of sequence sharing is greater across socio-cultural groups, than within groups. One possible explanation of this is that the separation into different ethnic groups is a relatively recent phenomenon. The nature and extent of sequence sharing across linguistic groups is even more interesting. The highest number of sequences are shared between Indo-European and Dravidian speakers. This is surprising because presently there is a clear separation of habitat between Indo-European (northern India) and Dravidian (southern India) speakers. A parsimonious explanation of this finding is that the Dravidian speakers were widely present in northern India prior to the arrival of the Indo-European speakers and there was considerable admixture before the Dravidian speakers retreated to occupy the southern regions of India. This explanation is consistent with the views of some anthropologists (Sarkar 1958). It is also interesting that the extent of sequence sharing between Austro-Asiatics and the other linguistic groups is much smaller, indicating that the Austro-Asiatics may have remained isolated for a considerable period of time. Since there is

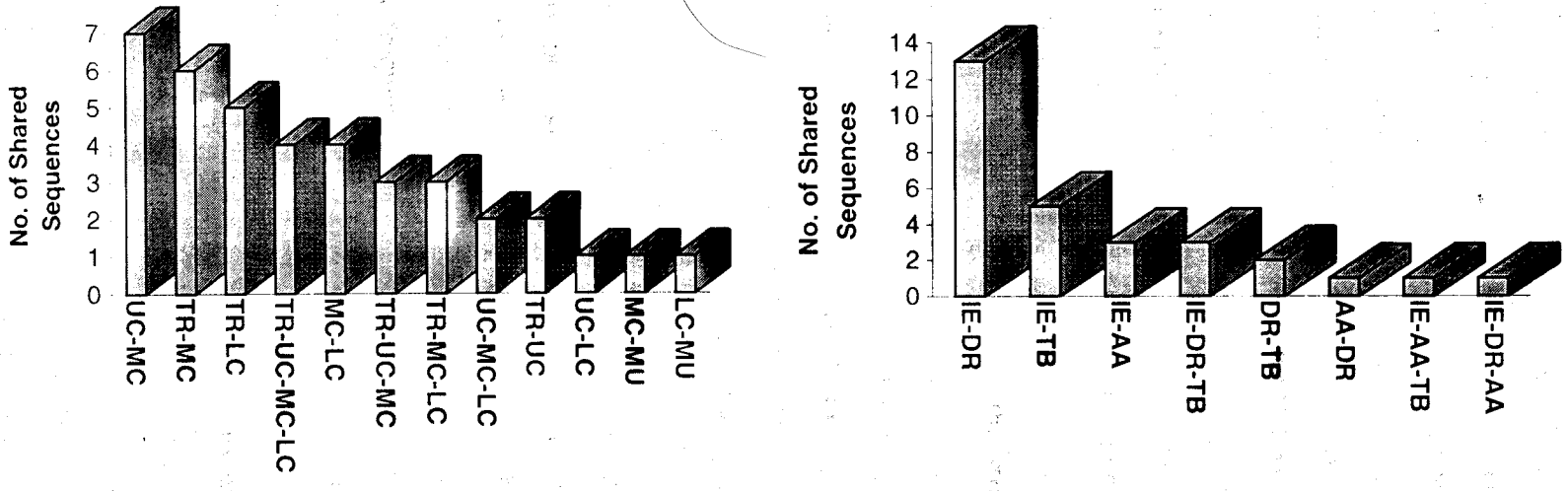



**Figure 2.5**
Numbers of mitochondrial HVS1 sequences shared among different subgroups of populations. [UC, Upper Caste; MC, Middle Caste; LC, Lower Caste; TR, Tribal; MU, Muslim; IE, AA, Austro-Asiatic; DR, Dravidian; IE, Indo-European; TB, Tibeto-Burman]

considerable overlap/confounding of linguistic and geographical groupings, results of sequence sharing across geographical regions are similar to those described with respect to language groups and are, therefore, not separately presented.

In order to further analyze the nature of sequence variation, we grouped individuals into M and U haplogroups. It was found that 131 of the 153 variable positions are also variable among individuals (n=338) belonging to HG-M, while among individuals (n=115) belonging to HG-U only 73 positions are variable (Table 2.4). To investigate existence of possible motifs (**descriptions of the probabilistic search algorithms used for motif finding are given in Chapter 3**) within HG-M and within HG-U, we selected those nucleotide positions at which the total frequency of variant nucleotides exceeded $\alpha$ proportion of the total sample size. For HG-M, $\alpha$ was chosen as 0.05. For HG-U, $\alpha$ was chosen as 0.1, in view of the smaller total number of individuals belonging to HG-U. Based on this strategy, 19 sites were identified in respect of HG-M and 12 sites in respect of HG-U. These 19 sites yielded a total of 72 haplotypes among HG-M individuals, and the 12 sites yielded a total of 29 haplotypes among HG-U individuals. Among HG-M individuals, 98.22% possessed T at np 16223. Thus, the vast majority of HG-M individuals belong to the subhaplogroup M*, as was also found by Bamshad et al. (2001). When the frequencies of variant nucleotides at these sites were cross-tabulated by population, socio-cultural category, geographical region and linguistic group, no clear pattern emerged. The only exception is that among the Kota and the Toto, the vast majority of individuals (70% and 57%, respectively) possessed only two 19-site haplotypes, although these haplotypes are not shared between the two populations. We searched for the possible motifs, including those that have been reported earlier in the literature and used to define various subhaplogroups (Bamshad et al. 2001). The frequencies of the known subhaplogroups of M are presented in Table 2.5. Clade M2 is the most frequent one in Indian populations. It occurs in significantly ($p<.05$) higher frequencies among tribals (26.8%), particularly among the Austro-Asiatic tribals (32%), than among castes (8.8%). The nucleotide diversity within this clade is higher than the other major clades, suggesting that this clade is ancient in India. The high frequency of this clade in Austro-Asiatic tribals supports our contention that they may be the earliest settlers in India. Although there are differences in frequencies of the major clades among social/geographical groups, these frequecies do not provide any significant discrimination between groups. In other words, while these clades may have been brought into India at different points of time, subsequent admixture

**Table 2.4**
**Frequencies of Observed Nucleotides at the Various Polymorphic Positions in the Pooled Sample as well as in Samples on Haplogroup M and Haplogroup U Backgrounds**

| Sl. No. | np - 16000 | nt in CRS | Frequency of | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | | | T | | | G | | | C | | | Deletion | | |
| | | | Pooled | M | U | Pooled | M | U | Pooled | M | U | Pooled | M | U | Pooled | M | U |
| 1 | 37 | A | 525 | 335 | 115 | | | | 3 | 3 | | | | | | | |
| 2 | 38 | A | 525 | 335 | 115 | | | | 3 | 3 | | | | | | | |
| 3 | 48 | G | 5 | 5 | | | | | 523 | 333 | 115 | | | | | | |
| 4 | 51 | A | 439 | 329 | 36 | | | | 89 | 9 | 79 | | | | | | |
| 5 | 66 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 6 | 69 | C | | | | 2 | | 2 | | | | 526 | 338 | 113 | | | |
| 7 | 75 | T | | | | 526 | 337 | 115 | | | | 2 | 1 | | | | |
| 8 | 81 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 9 | 82 | C | | | | 2 | 1 | 1 | | | | 526 | 337 | 114 | | | |
| 10 | 86 | T | | | | 503 | 315 | 113 | | | | 25 | 23 | 2 | | | |
| 11 | 92 | T | | | | 510 | 334 | 106 | | | | 18 | 4 | 9 | | | |
| 12 | 93 | T | 4 | | 4 | 492 | 314 | 108 | | | | 32 | 24 | 3 | | | |
| 13 | 94 | T | | | | 527 | 337 | 115 | | | | 1 | 1 | | | | |
| 14 | 95 | C | | | | 2 | 2 | | | | | 526 | 336 | 115 | | | |
| 15 | 102 | T | | | | 527 | 338 | 115 | | | | 1 | | | | | |
| 16 | 104 | C | 1 | | | 1 | 1 | | | | | 526 | 337 | 115 | | | |
| 17 | 108 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 18 | 110 | G | 1 | | 1 | | | | 527 | 338 | 114 | | | | | | |
| 19 | 111 | C | | | | 9 | 5 | | | | | 519 | 333 | 115 | | | |
| 20 | 124 | T | | | | 525 | 335 | 115 | | | | 3 | 3 | | | | |
| 21 | 126 | T | | | | 492 | 305 | 114 | | | | 36 | 33 | 1 | | | |
| 22 | 129 | G | 68 | 51 | 3 | | | | 454 | 287 | 106 | 6 | | 6 | | | |
| 23 | 134 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 24 | 136 | T | | | | 527 | 337 | 115 | | | | 1 | 1 | | | | |
| 25 | 140 | T | | | | 510 | 321 | 115 | | | | 18 | 17 | | | | |
| 26 | 142 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 27 | 144 | T | 4 | 4 | | 524 | 334 | 115 | | | | | | | | | |
| 28 | 145 | G | 20 | 12 | 4 | | | | 508 | 326 | 111 | | | | | | |
| 29 | 147 | C | | | | 3 | 1 | | 1 | 1 | | 524 | 336 | 115 | | | |
| 30 | 148 | C | | | | 2 | 2 | | | | | 526 | 336 | 115 | | | |
| 31 | 153 | G | 4 | 1 | | | | | 524 | 337 | 115 | | | | | | |
| 32 | 154 | T | | | | 521 | 338 | 110 | | | | 7 | | 5 | | | |
| 33 | 156 | G | 1 | 1 | | | | | 527 | 337 | 115 | | | | | | |
| 34 | 158 | A | 523 | 333 | 115 | | | | 5 | 5 | | | | | | | |
| 35 | 162 | A | 525 | 338 | 115 | | | | 3 | | | | | | | | |
| 36 | 163 | A | 527 | 338 | 114 | | | | 1 | | 1 | | | | | | |
| 37 | 164 | A | 527 | 338 | 115 | | | | 1 | | | | | | | | |
| 38 | 166 | A | 525 | 335 | 115 | | | | | | | | | | 3 | 3 | |

... continued

**Table 2.4 … continued**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 39 | 167 | C | | | | 3 | 2 | 1 | | | | 525 | 336 | 114 | | | |
| 40 | 168 | C | | | | 11 | 1 | 8 | | | | 517 | 337 | 107 | | | |
| 41 | 169 | C | | | | 5 | 1 | 4 | | | | 523 | 337 | 111 | | | |
| 42 | 172 | T | | | | 504 | 331 | 110 | | | | 24 | 7 | 5 | | | |
| 43 | 173 | C | | | | 2 | | 2 | | | | 526 | 338 | 113 | | | |
| 44 | 176 | C | | | | 12 | 12 | | | | | 516 | 326 | 115 | | | |
| 45 | 178 | T | | | | 525 | 338 | 114 | | | | 3 | | 1 | | | |
| 46 | 179 | C | | | | 14 | 11 | 3 | | | | 514 | 327 | 112 | | | |
| 47 | 180 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 48 | 181 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 49 | 182 | A | 517 | 332 | 111 | | | | | | | | | | 11 | 6 | 4 |
| 50 | 183 | A | 479 | 307 | 106 | | | | 1 | 1 | | 2 | | | 46 | 30 | 9 |
| 51 | 184 | C | | | | 10 | 9 | 1 | | | | 518 | 329 | 114 | | | |
| 52 | 185 | C | | | | 10 | 7 | | | | | 518 | 331 | 115 | | | |
| 53 | 186 | C | | | | 3 | 3 | | | | | 525 | 335 | 115 | | | |
| 54 | 187 | C | | | | 5 | 3 | 1 | | | | 523 | 335 | 114 | | | |
| 55 | 188 | C | | | | 6 | 5 | | | | | 522 | 333 | 115 | | | |
| 56 | 189 | T | | | | 445 | 285 | 99 | | | | 82 | 52 | 16 | 1 | 1 | |
| 57 | 190 | C | | | | | | | | | | 526 | 338 | 115 | 2 | | |
| 58 | 192 | C | | | | 14 | 9 | 2 | | | | 514 | 329 | 113 | | | |
| 59 | 193 | C | | | | 4 | 3 | 1 | | | | 524 | 335 | 114 | | | |
| 60 | 206 | A | 504 | 338 | 91 | | | | | | | 24 | | 24 | | | |
| 61 | 207 | A | 520 | 338 | 107 | | | | 8 | | 8 | | | | | | |
| 62 | 209 | T | | | | 513 | 333 | 105 | | | | 15 | 5 | 10 | | | |
| 63 | 213 | G | 5 | 3 | 1 | | | | 523 | 335 | 114 | | | | | | |
| 64 | 214 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 65 | 215 | A | 526 | 337 | 114 | | | | 2 | 1 | 1 | | | | | | |
| 66 | 217 | T | | | | 523 | 337 | 115 | | | | 5 | 1 | | | | |
| 67 | 218 | C | | | | 6 | 4 | 2 | | | | 522 | 334 | 113 | | | |
| 68 | 220 | A | 527 | 338 | 114 | | | | | | | 1 | | 1 | | | |
| 69 | 222 | C | | | | 1 | | | | | | 527 | 338 | 115 | | | |
| 70 | 223 | C | | | | 351 | 332 | 6 | | | | 177 | 6 | 109 | | | |
| 71 | 224 | T | | | | 521 | 335 | 112 | | | | 7 | 3 | 3 | | | |
| 72 | 225 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 73 | 227 | A | 526 | 338 | 114 | | | | 2 | | 1 | | | | | | |
| 74 | 230 | A | 508 | 336 | 97 | | | | 20 | 2 | 18 | | | | | | |
| 75 | 231 | T | | | | 524 | 334 | 115 | | | | 4 | 4 | | | | |
| 76 | 234 | C | | | | 37 | 17 | 20 | | | | 491 | 321 | 95 | | | |
| 77 | 235 | A | 526 | 338 | 115 | | | | 2 | | | | | | | | |
| 78 | 239 | C | | | | 17 | 1 | 12 | | | | 511 | 337 | 103 | | | |
| 79 | 240 | A | 524 | 335 | 114 | | | | 3 | 3 | | 1 | | 1 | | | |
| 80 | 241 | A | 527 | 338 | 115 | | | | 1 | | | | | | | | |
| 81 | 242 | C | | | | 3 | 3 | | 1 | | 1 | 524 | 335 | 114 | | | |
| 82 | 243 | T | | | | 519 | 329 | 115 | | | | 9 | 9 | | | | |
| 83 | 245 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |

… continued

**Table 2.4 … continued**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 84 | 246 | A | 526 | 336 | 115 | 2 | 2 | | | | | | | | | | |
| 85 | 247 | A | 521 | 338 | 108 | | | | 7 | | 7 | | | | | | |
| 86 | 248 | C | | | | 1 | | | | | | 527 | 338 | 115 | | | |
| 87 | 249 | T | | | | 520 | 336 | 110 | 1 | | | 7 | 2 | 5 | | | |
| 88 | 250 | C | | | | 1 | | | | | | 527 | 338 | 115 | | | |
| 89 | 254 | A | 526 | 338 | 113 | | | | 2 | | 2 | | | | | | |
| 90 | 256 | C | | | | 16 | 11 | 4 | | | | 512 | 327 | 111 | | | |
| 91 | 257 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 92 | 258 | A | 526 | 336 | 115 | | | | | | | | | | 2 | 2 | |
| 93 | 259 | C | 1 | 1 | | 1 | 1 | | | | | 526 | 336 | 115 | | | |
| 94 | 260 | C | | | | 8 | 2 | 2 | | | | 520 | 336 | 113 | | | |
| 95 | 261 | C | | | | 12 | 4 | 3 | | | | 516 | 334 | 112 | | | |
| 96 | 263 | T | | | | 526 | 336 | 115 | | | | 2 | 2 | | | | |
| 97 | 264 | C | | | | 3 | 3 | | | | | 525 | 335 | 115 | | | |
| 98 | 265 | A | 521 | 335 | 111 | | | | 1 | | 1 | 6 | 3 | 3 | | | |
| 99 | 266 | C | | | | 21 | 4 | 5 | | | | 507 | 334 | 110 | | | |
| 100 | 269 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 101 | 270 | C | | | | 26 | 23 | 2 | | | | 502 | 315 | 113 | | | |
| 102 | 272 | A | 520 | 330 | 115 | | | | 8 | 8 | | | | | | | |
| 103 | 274 | G | 41 | 35 | 4 | | | | 487 | 303 | 111 | | | | | | |
| 104 | 275 | A | 524 | 334 | 115 | | | | 4 | 4 | | | | | | | |
| 105 | 276 | T | | | | 526 | 336 | 115 | | | | 2 | 2 | | | | |
| 106 | 278 | C | | | | 17 | 11 | 5 | | | | 511 | 327 | 110 | | | |
| 17 | 284 | A | 524 | 335 | 115 | | | | 4 | 3 | | | | | | | |
| 108 | 286 | C | | | | 1 | | | | | | 527 | 338 | 115 | | | |
| 109 | 287 | C | | | | 2 | 2 | | | | | 526 | 336 | 115 | | | |
| 110 | 288 | T | | | | 527 | 338 | 115 | | | | 1 | | | | | |
| 111 | 289 | A | 512 | 323 | 115 | | | | 16 | 15 | | | | | | | |
| 112 | 290 | C | | | | 6 | | 2 | | | | 522 | 338 | 113 | | | |
| 113 | 291 | C | 2 | 1 | | 23 | 11 | 12 | | | | 503 | 326 | 103 | | | |
| 114 | 292 | C | | | | 6 | 1 | 1 | | | | 522 | 337 | 114 | | | |
| 115 | 293 | A | 525 | 336 | 115 | | | | 1 | 1 | | 2 | 1 | | | | |
| 116 | 294 | C | | | | 17 | 15 | 1 | | | | 511 | 323 | 114 | | | |
| 117 | 295 | C | | | | 17 | 15 | 2 | 3 | 3 | | 508 | 320 | 113 | | | |
| 118 | 296 | C | | | | 3 | 1 | | | | | 525 | 337 | 115 | | | |
| 119 | 297 | T | | | | 526 | 336 | 115 | | | | 2 | 2 | | | | |
| 120 | 298 | T | | | | 516 | 328 | 115 | | | | 12 | 10 | | | | |
| 121 | 299 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 122 | 300 | A | 521 | 333 | 114 | | | | 7 | 5 | 1 | | | | | | |
| 123 | 301 | C | | | | 1 | 1 | | | | | 527 | 337 | 115 | | | |
| 124 | 302 | A | 527 | 338 | 115 | | | | 1 | | | | | | | | |
| 125 | 304 | T | | | | 488 | 327 | 109 | | | | 40 | 11 | 6 | | | |
| 126 | 305 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 127 | 309 | A | 501 | 335 | 98 | | | | 27 | 3 | 17 | | | | | | |

… continued

**Table 2.4 ... continued**

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 128 | 311 | T | | | | 431 | 286 | 86 | | | | 97 | 52 | 29 | | | |
| 129 | 312 | A | 527 | 337 | 115 | | | | 1 | 1 | | | | | | | |
| 130 | 316 | A | 522 | 332 | 115 | | | | 6 | 6 | | | | | | | |
| 131 | 318 | A | 489 | 330 | 90 | 27 | 7 | 18 | 6 | | 3 | 6 | 1 | 4 | | | |
| 132 | 319 | G | 82 | 72 | 3 | | | | 446 | 266 | 112 | | | | | | |
| 133 | 320 | C | | | | 28 | 24 | 1 | | | | 500 | 314 | 114 | | | |
| 134 | 321 | C | | | | 3 | 3 | | | | | 525 | 335 | 115 | | | |
| 135 | 322 | A | 527 | 337 | 115 | 1 | 1 | | | | | | | | | | |
| 136 | 324 | T | | | | 523 | 333 | 115 | | | | 5 | 5 | | | | |
| 137 | 325 | T | | | | 512 | 333 | 115 | | | | 16 | 5 | | | | |
| 138 | 327 | C | | | | 6 | 6 | | | | | 522 | 332 | 115 | | | |
| 139 | 330 | T | | | | 527 | 337 | 115 | | | | 1 | 1 | | | | |
| 140 | 335 | A | 527 | 338 | 114 | | | | 1 | | 1 | | | | | | |
| 141 | 342 | T | | | | 527 | 337 | 115 | | | | 1 | 1 | | | | |
| 142 | 343 | A | 527 | 338 | 115 | | | | 1 | | | | | | | | |
| 143 | 344 | C | | | | 5 | 5 | | | | | 523 | 333 | 115 | | | |
| 144 | 352 | T | | | | 494 | 315 | 105 | | | | 34 | 23 | 10 | | | |
| 145 | 353 | C | | | | 13 | | 12 | | | | 515 | 338 | 103 | | | |
| 146 | 354 | C | 1 | | | | | | | | | 527 | 338 | 115 | | | |
| 147 | 355 | C | | | | 5 | 3 | 1 | | | | 523 | 335 | 114 | | | |
| 148 | 356 | T | | | | 510 | 330 | 111 | | | | 18 | 8 | 4 | | | |
| 149 | 357 | T | | | | 525 | 335 | 115 | | | | 3 | 3 | | | | |
| 150 | 359 | T | | | | 524 | 335 | 114 | | | | 4 | 3 | 1 | | | |
| 151 | 360 | C | | | | 3 | 1 | | | | | 525 | 337 | 115 | | | |
| 152 | 362 | T | | | | 436 | 277 | 103 | | | | 92 | 61 | 12 | | | |
| 153 | 368 | T | | | | 512 | 322 | 115 | | | | 16 | 16 | | | | |

**Table 2.5**
**Frequencies of Various known Sub-haplogroups of Haplogroups M and U in 44 Ethnic Populations of India and in Different Social, Geographical and Linguistic Groups**

| Variable | Code | M | | | | | | | | | U | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M* | M2 | M3 | M4 | M5 | M6 | M4a | M7a | M7b | U1 | U2i | U2e | U4 | U5 | U7 |
| Popula- | AGH | 5 | 1 | 3 | | 1 | | | | | | 4 | | | | |
| tion | AMB | 8 | 3 | | 5 | | | | | | | 2 | | | | |
| | BAG | 6 | | | 1 | 2 | | | | | | 3 | | | | 1 |
| | CHA | 4 | | | 2 | | 1 | | | | | 3 | | | | 1 |
| | CHK | 3 | | | | | | | | | | | | | | |
| | GAU | 6 | 4 | 1 | | | | | | | | 2 | | | | |
| | GND | 4 | 2 | 1 | | | | | | | | 4 | | | | |
| | HAL | 12 | 5 | | | 4 | | | | | | 3 | | | | |
| | HO | 6 | 5 | | | | 1 | | | | | 4 | | | | |
| | ILA | 5 | | | | | | | | | | 7 | | | | |
| | IYN | 7 | | 1 | | 1 | | | | | | 3 | | | | |
| | IYR | 5 | | 1 | 1 | 1 | | | | | 1 | 2 | 1 | | | 1 |
| | JAM | 9 | | | 1 | 2 | | | | 1 | | | | | | 1 |
| | JSK | 8 | | | | 3 | 1 | | | | 1 | | | | 1 | |
| | KBR | 8 | 6 | | 2 | | | | | | | | | | | 1 |
| | KHT | 10 | 1 | 2 | 1 | 2 | | | | | | | | | | |
| | KMR | 6 | 2 | | | 1 | | | | | | 4 | | | | |
| | KOT | 24 | | 12 | 2 | | | | | | | | | | | |
| | KUR | 8 | 3 | | | 1 | | | | | 2 | | | | | |
| | LOD | 9 | 1 | 3 | | 2 | | | | | | 5 | | | | 1 |
| | MAH | 5 | | 1 | 2 | 1 | | | | | | 5 | | | | |
| | MNP | 5 | | | | 1 | | | | | | 1 | | | | |
| | MOG | 10 | | | 1 | 5 | | | 5 | | | | | | | |
| | MRT | 5 | 1 | | 4 | 1 | | | | | | 2 | 1 | | | 1 |
| | MUN | 5 | 2 | | | | | | | | | | | | | |
| | MUR | 6 | 1 | | 2 | 2 | | | | | | 2 | | | | |
| | MUS | 5 | | | 2 | | | | | | | | | | | 2 |
| | MZO | 10 | 1 | | 4 | | | | | | | | | | | |
| | NBH | 7 | 2 | 1 | 3 | 1 | | | | | | 3 | 1 | | | |
| | PBR | | | | | | | | | | | | | | | 1 |
| | PLN | 8 | 1 | | 1 | 3 | | | | | | 2 | | | | |
| | RAJ | 5 | | 1 | 1 | 1 | | | | | | 4 | 1 | | | |
| | RIA | 10 | 2 | 2 | | | | | | | | | | | | |
| | SAN | 8 | 1 | | | 4 | | | | | | 2 | | | | |
| | SBR | 11 | 2 | | 2 | 2 | | | | | | 2 | | | 1 | |
| | SCH | 10 | 1 | 1 | 1 | 2 | | | | | | | | | | 2 |
| | TAN | 8 | 3 | 1 | | 2 | | | | | 1 | 1 | | | | 1 |
| | TOD | 9 | 7 | | | | | | | | | | | | | |
| | TRI | 6 | 1 | | 2 | | | 1 | | | | 3 | | | | |
| | TTO | 18 | 12 | | 5 | | | | | | | | | | | |
| | UBR | 5 | | | 1 | | | | | | | 3 | 2 | | | |
| | VAN | 7 | | | | 2 | 1 | | | | | 1 | | | | |
| | VLR | 8 | | | 2 | | | | | | | 1 | | | | |
| | WBR | 8 | 2 | 1 | 2 | 2 | | | | 1 | | 1 | | | | |

... continued

**Table 2.5 ... continued**

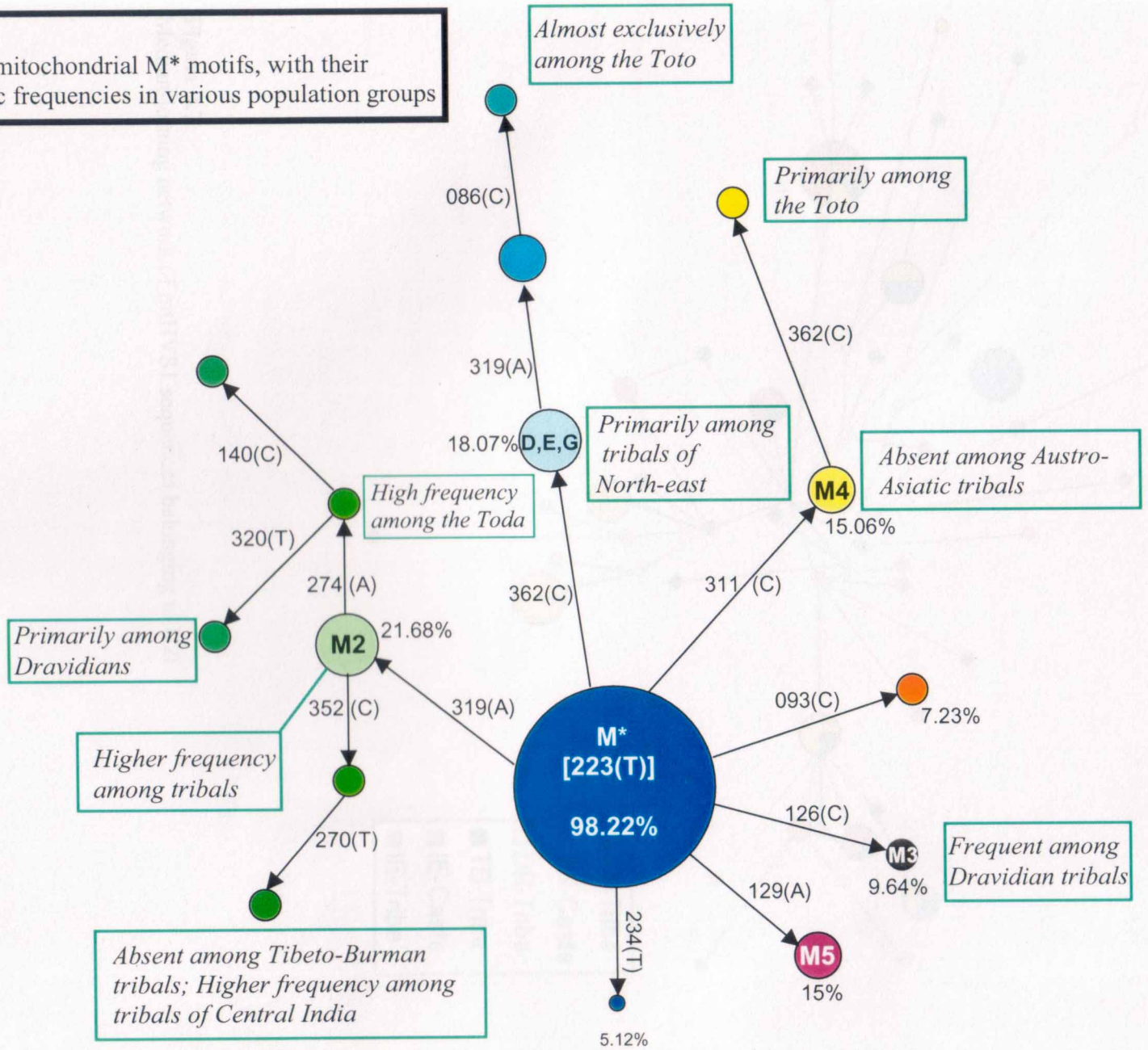
| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Socio-Cultural Group | TRIBE | 168 | 45 | 18 | 17 | 21 | 1 | 1 | 5 | 1 | 2 | 34 | | | | 2 |
| | OTHER | 5 | | | 2 | | | | | | | | | | | 2 |
| | UPPER CASTE | 49 | 10 | 3 | 8 | 7 | | | | 1 | 1 | 12 | 3 | | 1 | 3 |
| | MIDDLE CASTE | 67 | 10 | 8 | 15 | 11 | 2 | | | | 1 | 21 | 2 | | 1 | 1 |
| | LOWER CASTE | 43 | 7 | 3 | 8 | 10 | 1 | | | | 1 | 12 | 1 | | | 5 |
| Geographical Region | NORTH | 47 | 2 | 4 | 8 | 8 | 2 | | | | 1 | 10 | 3 | | 1 | 6 |
| | NORTH EAST | 71 | 16 | 2 | 13 | 8 | | 1 | 5 | 1 | | 2 | | | | 1 |
| | EAST | 66 | 19 | 10 | 5 | 14 | 1 | | | 1 | 1 | 27 | | | | 3 |
| | SOUTH | 89 | 14 | 14 | 11 | 8 | 1 | | | | 3 | 18 | 1 | | | 1 |
| | CENTRAL | 39 | 12 | 1 | 4 | 9 | | | | | | 15 | | | 1 | |
| | WEST | 20 | 9 | 1 | 9 | 2 | | | | | | 5 | 2 | | | 2 |
| Linguistic Group | AA | 28 | 9 | 3 | | 6 | 1 | | | | | 11 | | | | 1 |
| | DR | 105 | 19 | 15 | 13 | 11 | 1 | | | | 3 | 28 | 1 | | | 1 |
| | TB | 71 | 16 | 2 | 13 | 8 | | 1 | 5 | 1 | | 4 | | | | 1 |
| | IE | 128 | 28 | 12 | 24 | 24 | 2 | | | 1 | 2 | 36 | 5 | | 2 | 10 |

and/or formations of new groups through fission of existing groups have resulted in diffusion of clades across populations and geographical space. We have, however, discovered various motifs, some of which have not been reported earlier in the literature. The evolutionary network of these motifs, based on the network analysis of Bandelt et al. (1999), with considerable simplifications and culling for the purpose of presentation of its major features, is given in Figure 2.6. Some of these motifs occur exclusively or in high frequencies in some populations/groups/regions, or are absent from some regions/groups. These data are also provided in Figure 2.6. This figure indicates that several motifs are concentrated in specific ethnic groups, consistent with fission and founder effect.

Among individuals belonging to HG-U, the frequencies of known subhaplogroups are also presented in Table 2.5. HG-U is a complex mtDNA lineage, with an estimated age of 51000-67000 years (Torroni et al. 1996). Our estimate of about 45000 ± 25,000 years (Table 2.6) is not significantly different from the earlier estimate. The vast majority of HG-U individuals belong to HG-U2i and U7. The U2i is the Indian-specific subcluster of U (as opposed to the western-Eurasian subcluster U2e) and the coalescence age of the U7 subcluster in India has been estimated to be 32000 ± 5500 years (Kivisild et al. 1999a). Most tribal groups in India possess high frequencies of U2i, but not U7 (Table 2.5). On the other hand, the Indo-European speakers in India possess high frequencies of both U2i and U7 (Table 2.5). This indicates, as has already been suggested by Kivisild et al. (1999a), that the U2i predates the arrival of the Indo-European speakers from Central and West Asia into India. Interestingly, a motif GCGC at nps 16051, 16206, 26230 and 16311, respectively, has been found on the U2i background, which is present in 18 (16%) of the 115 individuals. This motif is found almost exclusively among tribal, middle- and lower-caste populations, but not among the upper-caste populations or the Muslims (of Uttar Pradesh). This motif is also detectable among many of the Pakistani samples screened by Kivisild et al. (1999a, see supplementary material).

Since U2i appears to be the indigenous subHG of U in India, we sought to find phylogenetic relationships among the distinct HVS1 sequences within this subHG and their observed frequencies in various linguistic groups. The phylogenetic network (Bandelt et al. 1999) is complex (Figure 2.7). There is no major starlike cluster to indicate sudden population exapansions nor is there any clear socio-linguistic clustering of related sequences.

**Figure 2.6**
Network of mitochondrial M* motifs, with their characteristic frequencies in various population groups



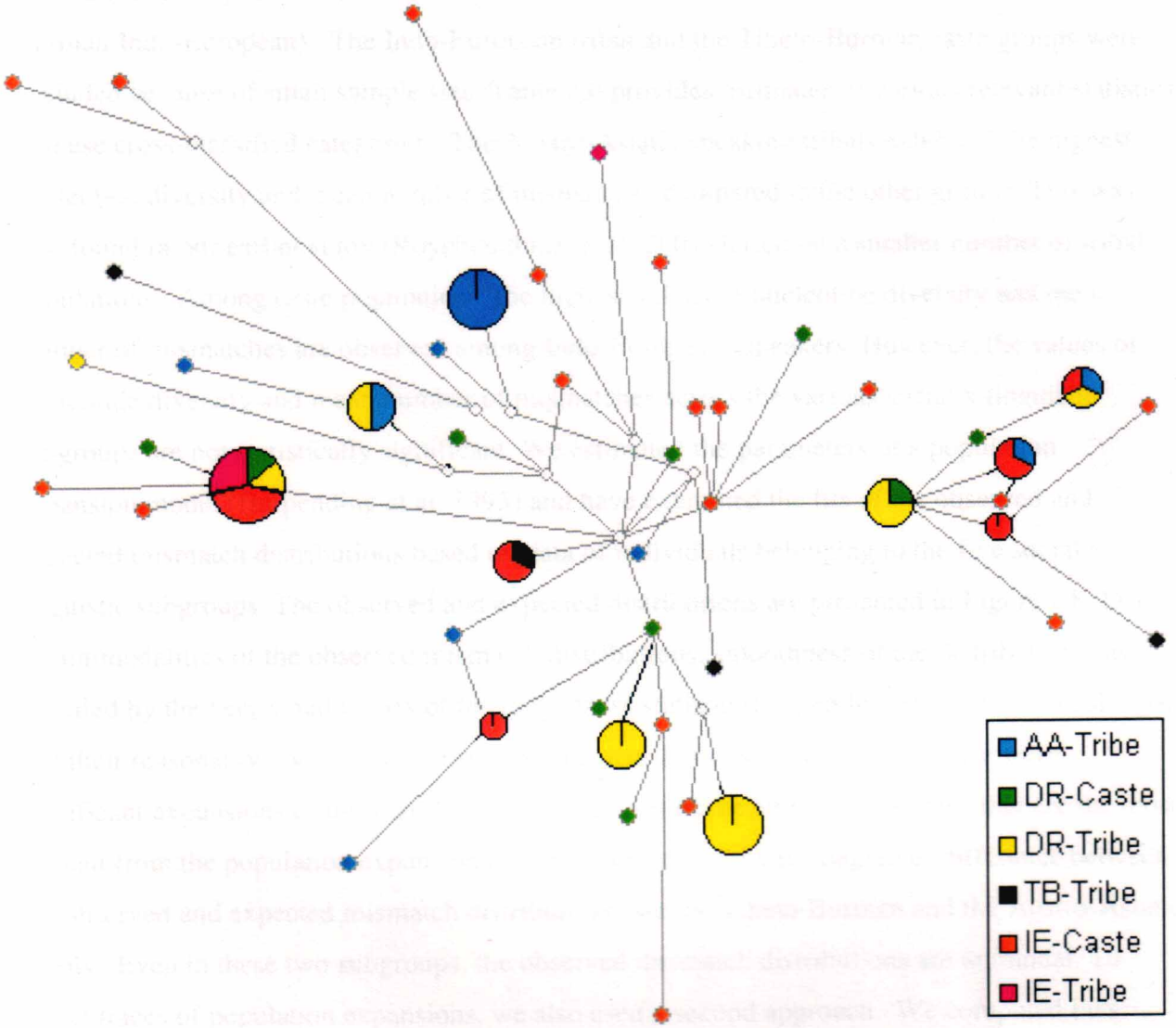



**Figure 2.7**
Median-joining network of mtHVS1 sequences belonging to U2i

Nucleotide Diversity and Mismatch Statistics: To avoid vagaries of small sample sizes, nucleotide diversities and mismatch statistics were calculated by pooling samples. Samples were cross-classified by social (tribe/caste) and linguistic affiliation (Austro-Asiatic/Dravidian/Tibeto-Burman/Indo-European). The Indo-European tribal and the Tibeto-Burman caste groups were excluded because of small sample size. Table 2.6 provides estimates of various relevant statistics in these cross-classified categories. The Austro-Asiatic speaking tribals exhibited the highest nucleotide diversity and mean number of mismatches compared to the other groups. This was also found in our earlier study (Roychoudhury et al. 2001) based on a smaller number of tribal populations. Among caste populations, the highest values of nucleotide diversity and mean number of mismatches are observed among Indo-Europeans speakers. However, the values of nucleotide diversity and mean number of mismatches across the various social x linguistic subgroups are not statistically significant. We estimated the parameters of a population expansion model (Harpending et al. 1993) and have examined the fits of the observed and expected mismatch distributions based on data of individuals belonging to the five social x linguistic subgroups. The observed and expected distributions are presented in Figure 2.8. From the unimodalities of the observed mismatch distributions, smoothness of the distributions [as revealed by the very small values of the raggedness statistic (Harpending et al. 1993); Table 2.6] and their reasonably good fits with the expected distributions, it is clear that there were significant expansions of these groups. The subgroups of the ethnic populations that are the most deviant from the population expansion model, as evidenced by the degree of difference between the observed and expected mismatch distributions, are the Tibeto-Burman and the Austro-Asiatic tribals. Even in these two subgroups, the observed mismatch distributions are unimodal. To detect traces of population expansions, we also used a second approach. We computed Fu's (1997) $F_S$ statistic, which is particularly sensitive to population growth. Significantly large negative values indicate population expansion (Fu 1997), which is what is observed (Table 2.6) in all social x linguistic subsets of individuals. We estimated the expansion times, which are also presented in Table 2.6. The estimated expansion time of the Austro-Asiatic tribals ($\approx$55,000 ybp) is much higher than that of the Dravidian or the Tibeto-Burman tribals ($\approx$42,000 ybp). However, the 95% confidence intervals of the expansion times is overlapping, indicating that the estimate expansion times of the three linguistic groups of tribals may not, however, be significantly different. The Dravidian and Indo-European castes do not show any significant

Table 2.6

**Numbers of Polymorphic Sites, Nucleotide Diversities, Mismatch Statistics, Estimated Expansion Times and other Statistics Pertaining to Population Expansion Based on HVS1 Sequence Data Classified by Social, Linguistic and Haplogroup Affiliation**

| | TRIBE | | | CASTE | | HAPLOGROUP | |
|---|---|---|---|---|---|---|---|
| | LINGUISTIC GROUP | | | LINGUISTIC GROUP | | | |
| | AUSTRO-ASIATIC | DRAVIDIAN | TIBETO-BURMAN | DRAVIDIAN | INDO-EUROPEAN | M | U |
| No. of Sequences | 46 | 94 | 95 | 60 | 195 | 338 | 115 |
| No. of Polymorphic Sites | 57 | 55 | 74 | 63 | 115 | 124 | 71 |
| Nucleotide Diversity ($\pi$)± SD | 0.0224 ± 0.0059 | 0.0170 ± 0.0045 | 0.0173 ± 0.0046 | 0.0154 ± 0.0042 | 0.0176 ± 0.0046 | 0.0150 ± 0.0040 | 0.0184 ± 0.0049 |
| Mean No. of Mismatches ± SD | 8.00 ± 1.89 | 6.07 ± 1.46 | 6.16 ± 1.48 | 5.51 ± 1.35 | 6.28 ± 1.50 | 5.37 ± 1.30 | 6.59 ± 1.57 |
| Expansion Time, yrs. (Range) | 54656 (40243-69024) | 41470 (30488-52439) | 42085 (30976-53415) | 37644 (27439-47683) | 42905 (31707-54146) | 36688 (26829-46341) | 45023 (33049-56707) |
| Raggedness, *r* | .0199 | .0069 | .0200 | .0094 | .0097 | .0161 | .0141 |
| Fu's $F_s$ (p-value) | -24.93 (0.0) | -25.26 (0.0) | -25.20 (0.0) | - 25.45 (0.0) | - 24.87 (0.0) | - 24.87 (0.0) | -25.03 (0.0) |

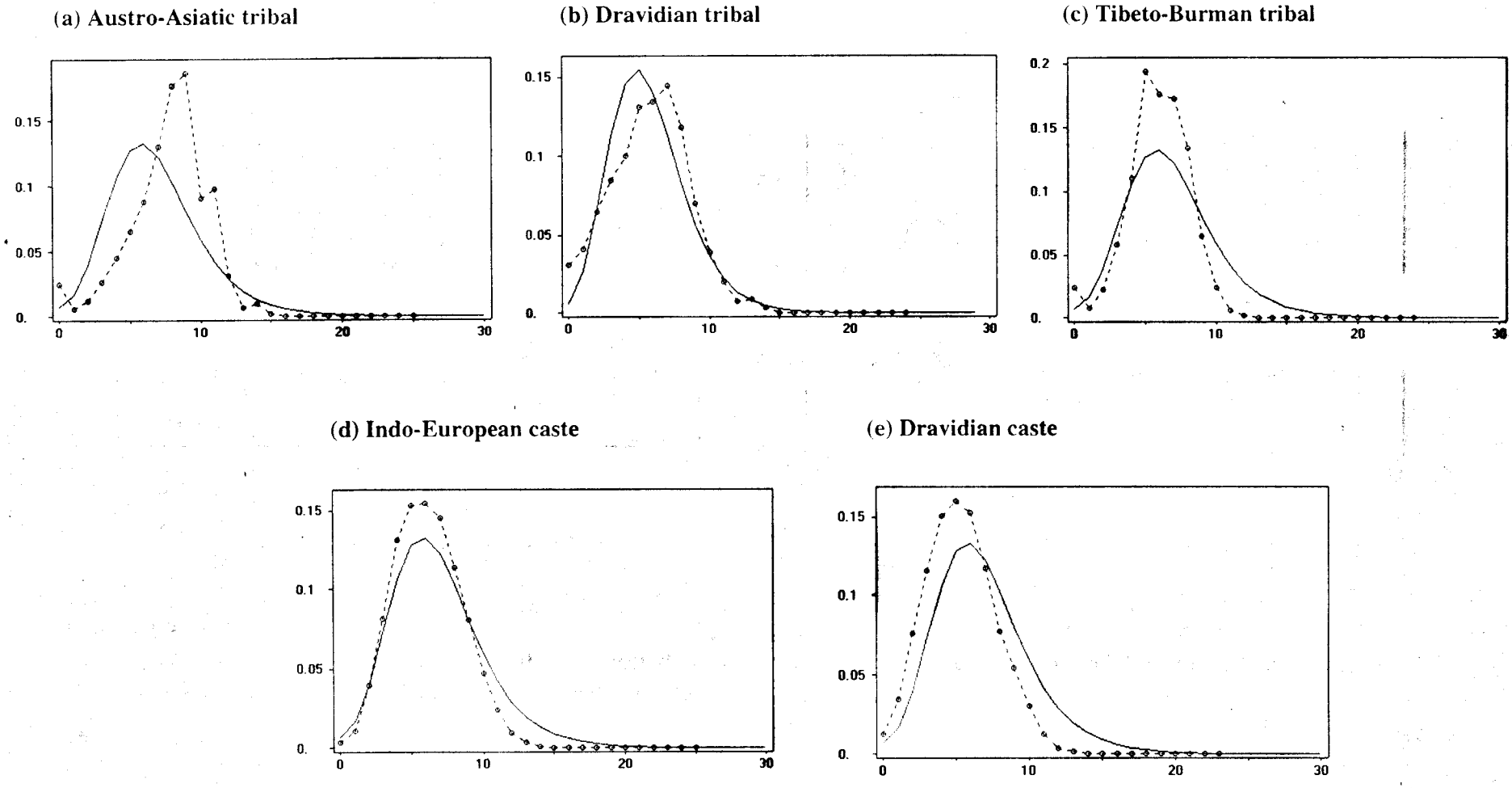
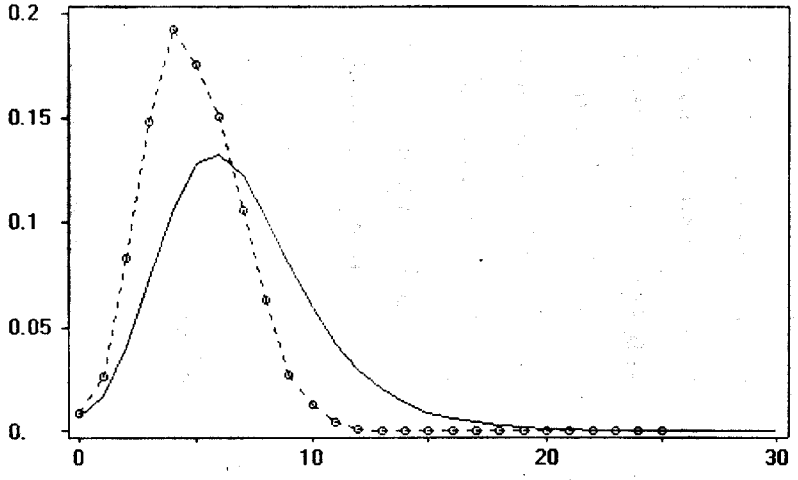



**Figure 2.8**
Observed (----) mismatch distributions and those expected (——) under a population expansion model based on data of mitochondrial HVS1 sequences from various linguistic subgroups of tribals and castes

(a) M haplogroup




(b) U haplogroup

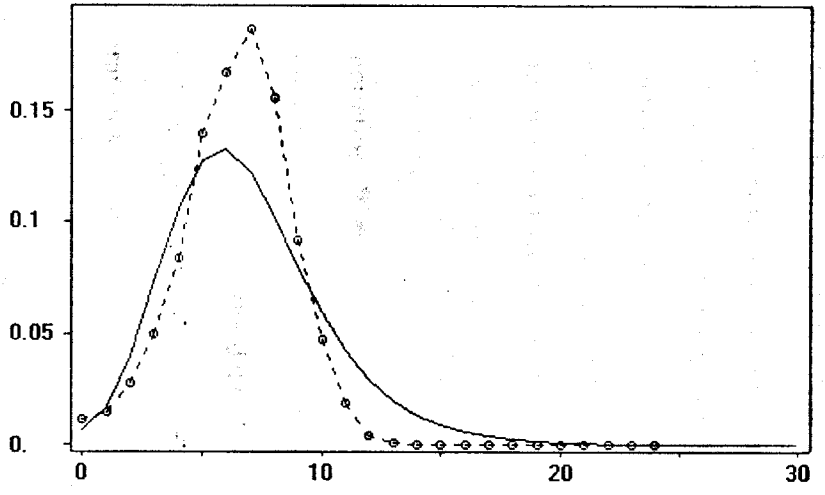



**Figure 2.9**
Observed (----) mismatch distributions and those expected ( —— ) under a population expansion model based on data of mitochondrial HVS1 sequences of individuals belonging to M and U haplogroups

difference in their estimated expansion times, and the estimates are similar to those of the Dravidian and Tibeto-Burman tribals. Thus, the expansion time of the Austro-Asiatic tribals is about 13,000 years earlier than of the other tribal and caste groups.

Nucleotide diversities and mean number of mismatches were also calculated among individuals belonging to HGs M and U (Table 2.6). Individuals belonging to HG-U possess higher values of nucleotide diversity and mean number of mismatches than individuals belonging to HG-M. The mismatch distributions are unimodal and fit very well with those expected under the population expansion model (Figure 2.9). The values of the $F_S$ statistic are also negative and large.

Genetic Differentiation: In order to examine the nature and extent of variation among populations grouped according to certain well-defined criteria, we have estimated $F_{ST}$ statistics and have carried out AMOVA using data of both RSP haplotypes and HVS1 sequences. The results are presented in Table 2.7. The $F_{ST}$ values are highest for tribal populations inhabiting different geographical regions or belonging to different linguistic groups. We recall that that there is some degree of confounding between geographical region of habitat and linguistic groups, as has been mentioned earlier. Genetic differentiation among the caste groups is substantially smaller than among the tribal groups. These results are expected because the tribal groups are ancient and have been isolated for a longer period of time than caste groups. Formation of the caste system is in itself a relative recent event. Among the caste groups, the upper castes of different geographical regions show the highest degree of differentiation.

The AMOVA results indicate that the extent of variation is the highest among individuals within population groups, both in terms of RSP haplotype frequencies and HVS1 sequences. Between 85% and 99% of the variation is attributable to between-individuals within-populations. The extent of variation in RSP haplotype frequencies among upper castes of different geographical regions is the highest (11.1%). This implies that there is stonger geographical substructuring of upper caste populations, compared to populations of other ranks. Similar results were, however, not obtained in respect of middle or lower castes or tribal populations. This possibly indicates that the upper caste groups of different geographical regions may not have had similar origins or may have undergone different types and levels of admixture. This is plausible because upper caste groups of northern India may have had much higher gene flow from central Asian populations compared to upper castes of southern India. Socio-cultural effects on mtDNA

Table 2.7
Estimates of $F_{ST}$ and AMOVA Results Based on Mitochondrial, Y-chromosomal and Autosomal Polymorphisms for Different Groupings of the Populations Studied

| GROUPING | $F_{ST}$ | | | | | % variation attributable to* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Between Groups | | | | | Between Populations within Groups | | | | |
| | mt | | Y | | Auto | mt | | Y | | Auto | mt | | Y | | Auto |
| | RSP | HVS1 | RSP | STRP | | RSP | HVS1 | RSP | STRP | | RSP | HVS1 | RSP | STRP | |
| 2 Groups: Caste and Tribe | .112 | .102 | .233 | .211 | .053 | 2.38 | 0.62 | 7.87 | 0.42 | 0.99 | 8.81 | 9.56 | 14.97 | 10.04 | 4.25 |
| 6 Groups: Geographical Regions | .108 | .099 | .241 | .229 | .052 | 4.32 | 1.01 | 11.29 | 3.16 | 1.88 | 6.51 | 8.99 | 12.49 | 8.1 | 3.32 |
| 4 Groups: Linguistic groups | .109 | .101 | .243 | .232 | .055 | 2.64 | 1.29 | 8.51 | 1.86 | 1.89 | 8.29 | 8.90 | 15.30 | 8.97 | 3.46 |
| 3 Groups: Ranked Castes – Upper, Middle and Lower | .056 | .033 | .144 | .119 | .020 | 0.37 | ≈0 | 5.39 | ≈0 | 0.11 | 5.23 | 3.83 | 8.38 | 4.67 | 1.78 |
| Upper castes of different geographical regions | .101 | .045 | .089 | .249 | .008 | 11.10 | 1.71 | 4.02 | 3.67 | 0.98 | ≈0 | 2.78 | 4.83 | 1.24 | ≈0 |
| Middle castes of different geographical regions | .053 | .047 | .050 | .115 | .022 | 2.46 | 0.66 | ≈0 | 0.65 | ≈0 | 2.88 | 4.05 | 10.51 | 3.88 | 2.29 |
| Lower castes of four different geographical regions | .010 | .016 | .192 | .121 | .023 | 0.84 | ≈0 | ≈0 | ≈0 | ≈0 | 0.16 | 1.83 | 32.24 | 9.17 | 4.77 |
| 4 Groups: Tribes of four different geographical regions (North-east, East, South and Central) | .138 | .158 | .239 | .272 | .069 | 1.88 | 1.90 | 5.25 | 1.37 | 2.12 | 11.94 | 13.90 | 18.70 | 10.03 | 4.47 |
| 4 Groups: Tribes of four different linguistic groups | .136 | .157 | .243 | .275 | .069 | 0.40 | 0.06 | 6.86 | 0.82 | 1.80 | 13.15 | 15.39 | 17.54 | 10.52 | 4.84 |

* The percentages of variation attributable to "between individuals within groups" are not shown. These are obtainable by subtracting from 100 the sum of the percentages of total variation attributable to the other two sources of variation which are shown.

substructuring seems minimal as the proportions of variance attributable to caste-tribal group differences, or linguistic differences are quite low.

Since Central Asia is supposed to have been a major contributor to the Indian gene pool, particularly to the north Indian gene pool, and the migrants had supposedly moved to India through Afghanistan and Pakistan, we have taken available data of HVS1 sequences from the Central Asian and Pakistani populations and have analyzed them jointly with the data generated during the present study. The data of the Central Asian region pertained to 103 sequences from Russia, Almaty region and Kazakhstan (Comas et al. 1998), 74 sequences from Turkey (Calafell et al. 1996) and 8 sequences from Pakistan (Kivisild et al. 1999a). We have computed a measure of genetic differentiation, $F_{ST}$, between the populations of the Central Asian and Pakistani regions and populations belonging to various geographical regions of India. These results are presented in Table 2.8, and revealed that the coefficient of genetic differentiation is the lowest between the Central Asian and Pakistani populations and the north Indian populations, higher between the south Indian populations and the highest between the northeast Indian populations. These results, therefore, reinforce the view that the extent of admixture between the Central Asian and northern Indian gene pools have been much higher than those with populations of other Indian regions.

Phylogenetic analysis: To examine the genetic affinities among the 44 ethnic populations a Neighbor Joining (NJ) tree was constructed using Nei's $D_A$ distance based on RSP haplotype frequencies. The unrooted NJ tree is presented in Figure 2.10. Although there is no strong clustering of populations belonging to the same social, geographical or linguistic group, two small clusters of populations – one comprising several populations of the north (top of the tree in Figure 2.10) and another comprising several populations of the northeast (bottom of the tree) are discernible. The lack of strong clustering is not surprising in view of the overwhelmingly strong sharing of haplotypes across populations noted earlier.

Of the 528 HVS1 sequences generated in the present study, 205 sequences were found to be shared by at least two individuals. Using the 323 distinct sequences, a neighbor-joining (NJ) tree was also constructed (figure not shown). There was again no clustering of the distinct sequences by geographical region, social rank or linguistic group. However, most of the sequences belonging to haplogroup U formed a separate clade.

**Figure 2.10**
Neighbor-joining tree depicting genetic affinities among Indian ethnic populations based on mitochondrial RSP frequencies

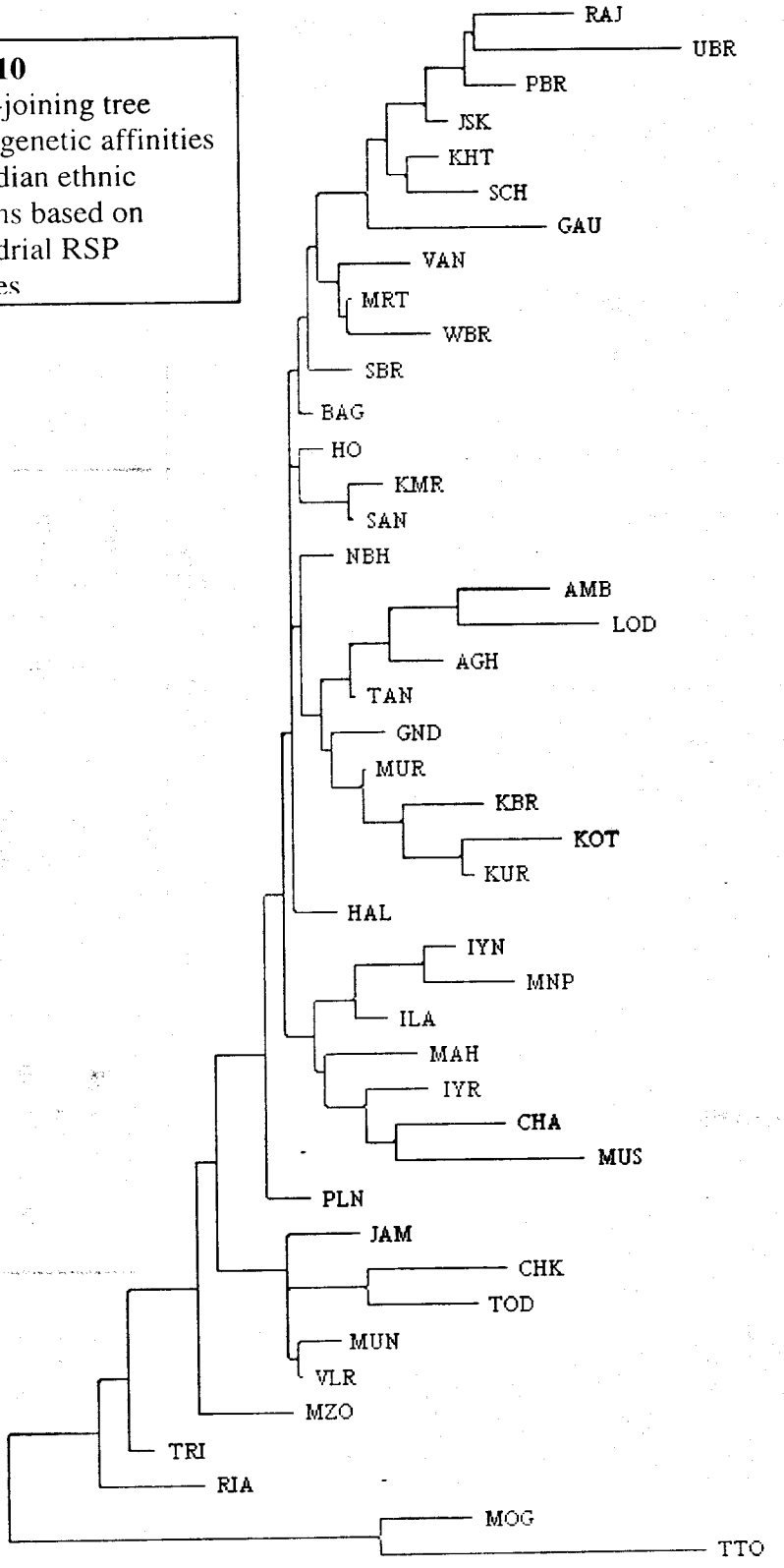


Table 2.8

**Genetic Differentiation, Estimated $F_{ST}$ values, between Various Subgroups of Indian Populations and Central Asian and Pakistani Populations based on HVS1 Sequence Data**

| North India | North Indian Upper Castes | North Indian Upper and Middle Castes | North Indian Upper and Middle Castes and Muslims | North-East India | North-Eastern Tribes | South India | South Indian Upper and Middle Castes |
|---|---|---|---|---|---|---|---|
| 0.017 | 0.016 | 0.012 | 0.015 | 0.043 | 0.047 | 0.042 | 0.039 |

We have also formally tested whether there is any association between geographic and genetic distances by using the nonparametric Mantel test. No statistically significant association was found, irrespective of the set of marker loci used for computing the genetic distances

*Y-chromosomal Polymorphisms*

Distribution of Y Haplogroups: On the basis of the RSP markers, we have classified Y chromosomes into six haplogroups (HGs) as defined by Rosser et al. (2000). The haplogroup frequencies in various cross-classified subgroups of populations are presented in Table 2.9, and their geographical distributions are presented in Figure 2.11 (a) and (b) separately for tribal and caste populations, respectively. As seen from Table 2.9, there are dramatic differences in the frequencies of the various haplogroups. First, the tribal populations, irrespective of linguistic affinity, possess significantly ($p<0.001$) lower frequencies of HG-1 compared to the caste populations. Both Dravidian speaking and Indo-European speaking caste populations show similar frequencies of both HG-1 ($\approx 20\%$) and HG-2 ($\approx 32\%$). HG-1 is probably arose in Central Asia (Zerjal et al. 2002). Therefore, it is not surprising that this haplogroup is more prevalent among the castes than among the tribals. However, contrary to expectations, HG-1 frequencies are higher among the Dravidian upper caste populations than among the Indo-European upper caste populations (Table 2.9). With respect to HG-2, there are statistically significant ($p<.0001$) differences in frequencies among Tibeto-Burman tribals ($\approx 10\%$), Austro-Asiatic tribals ($\approx 30\%$) and Dravidian speaking tribals ($\approx 70\%$). This haplogroup is the most ancestral lineage in Europe (Rosser et al. 2000), but this haplogroup probably contains a heterogeneous set of chromosomes that are not necessarily closely related (Zerjal et al. 2002). Dramatic differences in frequencies are also observed with respect to HG-26. The Tibeto-Burman tribals predominantly (70%) possess the HG-26. Similar high frequencies have also been observed (Su et al. 2000; Tajima et al. 2001) among the Han Chinese and several southeast Asian populations, which is consistent with the notion that the Han Chinese may have been the ancestral population from which the Tibeto-Burmans of India were derived. HG-28 is present among Dravidian and Indo-European speakers, but not among Austro-Asiatic and Tibeto-Burman speakers. This is again somewhat surprising since HG-28 is a derivative of HG-26. There are no significant differences in frequencies of HG-9 among the population subgroups. This haplogroup is commonly found in the Middle East (Hammer et al. 2000; Semino et al. 2000; Quintana-Murci et al. 2001),

**Table 2.9**
**Frequencies of Y Haplogroups in Ethnic Populations of India Belonging to Various Social, Linguistic and Geographical Groups**

| LINGUISTIC AFFILIATION/ GEOGRAPHICAL ZONE | SOCIAL GROUP | SAMPLE SIZE | HAPLOGROUP (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 9 | 26 | 28 |
| Austro-Asiatic | Tribe | 52 | 7.7 | 26.9 | 0.0 | 13.5 | 51.9 | 0.0 |
| Dravidian | Tribe | 84 | 4.8 | 64.3 | 3.6 | 9.5 | 13.1 | 8.2 |
| | Caste | 103 | 12.6 | 31.1 | 18.4 | 22.3 | 5.8 | 9.7 |
| Tibeto-Burman | Tribe | 87 | 4.7 | 8.0 | 0.0 | 14.9 | 72.4 | 0.0 |
| Indo-European | Tribe | 19 | 0.0 | 36.7 | 15.8 | 21.1 | 26.4 | 0.0 |
| | Caste | 122 | 12.3 | 32.0 | 33.6 | 11.5 | 8.2 | 2.4 |
| | | | | | | | | |
| North | Muslim | 19 | 15.8 | 0.0 | 57.9 | 10.5 | 15.8 | 0.0 |
| | Upper Caste | 52 | 11.5 | 25.0 | 36.5 | 19.2 | 1.9 | 5.8 |
| | Lower Caste | 18 | 11.1 | 44.4 | 44.4 | 0.0 | 0.0 | 0.0 |
| North-east | Tribe | 126 | 6.3 | 23.0 | 1.6 | 14.3 | 49.2 | 0.8 |
| East | Tribe | 52 | 7.7 | 26.9 | 0.0 | 13.5 | 51.9 | 0.0 |
| | Upper Caste | 13 | 7.7 | 7.7 | 76.9 | 0.0 | 7.7 | 0.0 |
| | Middle Caste | 22 | 13.6 | 40.9 | 9.1 | 9.1 | 27.3 | 0.0 |
| | Lower Caste | 17 | 17.6 | 47.0 | 11.8 | 5.9 | 17.6 | 0.0 |
| South | Tribe | 35 | 2.9 | 54.3 | 8.6 | 14.3 | 5.7 | 14.3 |
| | Upper Caste | 38 | 18.4 | 18.4 | 28.9 | 13.2 | 5.3 | 15.8 |
| | Middle Caste | 47 | 10.6 | 40.4 | 8.5 | 31.9 | 6.4 | 2.1 |
| | Lower Caste | 11 | 0.0 | 36.4 | 18.2 | 27.3 | 0.0 | 18.2 |
| Central | Tribe | 42 | 0.0 | 50.0 | 7.1 | 14.3 | 23.8 | 4.8 |

southeastern Europe and its highest frequencies are found in the Caucasus-Anatolia region (Rosser et al. 2000). With respect to the overall profiles of Y-haplogroup frequencies, there are statistically significant differences among both tribals ($p<.0001$) and castes ($p=.02$) belonging to the various linguistic groups. Similarly, the profiles among population subgroups are also statistically significant with respect to the northern ($p=.006$), eastern ($p<.0001$) and southern ($p<.00001$) zones. (Since samples from only tribals were available from the central and north-eastern zones, tests of significance are not relevant.) Figure 2.11(a) shows that among tribals, there are geographical clines of Y-haplogroup frequencies: HG-2 and HG-28 increase from north-eastern through central to southern India, while HG-26 shows a reverse trend. No clear geographical trend of haplogroup frequencies is discernible among castes of different ranks (Figure 2.11(b)).

Allele frequency distributions at STRP loci and STRP haplotype distributions: Table 2.10 provides the frequency distributions of repeat numbers in among tribal and caste populations, classified by linguistic affinity, at the 10 STRP loci screened. Since the sample size of the Muslim group was small, we have excluded them from this table. There are dramatic differences in the modal allele sizes and frequencies across the various subsets of our sample. At the DYS19 locus, while the allele containing 16 repeats is modal among Austro-Asiatic, Dravidian and Indo-European tribals, the 15-repeats allele is the modal one among Tibeto-Burman tribals. The 16-repeats allele is also the modal allele in Central Asia (Perez-Lezaun et al. 1999), but not in West Asia (Nebel et al. 2000) where the 14-repeats allele is the modal one. There are also notable differnces in allele frequencies among castes of different ranks, both for Dravidian and Indo-European speaking caste groups. Perez-Lezaun et al. (1999) reported the discovery of two new alleles – containing 10 and 18 repeats – among Central Asian populations at the DYS388 locus It is most interesting that the 10-repeats allele is the modal allele among Tibeto-Burman tribals of India. In all other populations, tribal and caste, the allele containing 12-repeats is the modal one. The 12-repeats allele is also the modal allele in both Central and West Asia (Perez-Lezaun et al. 1999; Nebel et al. 2000). Since the allele containing 10-repeats has low frequencies in Central Asia, it is possible that this allele was carried to Central Asia on a back migration from Asia to Africa (Hammer et al 1998; Cruciani et al. 2002). At the DYS389I locus, different Central Asian populations have different modal alleles; either 10- or 11-repeats. In the various subgroups of the Indian populations, this is generally true, except for the Tibeto-Burman tribals

**Figure 2.11**
Frequency distributions of Y-haplogroups in (a) tribal and (b) caste (UC, Upper Caste; MC, Middle Caste; LC, Lower Caste) populations of India

(a)

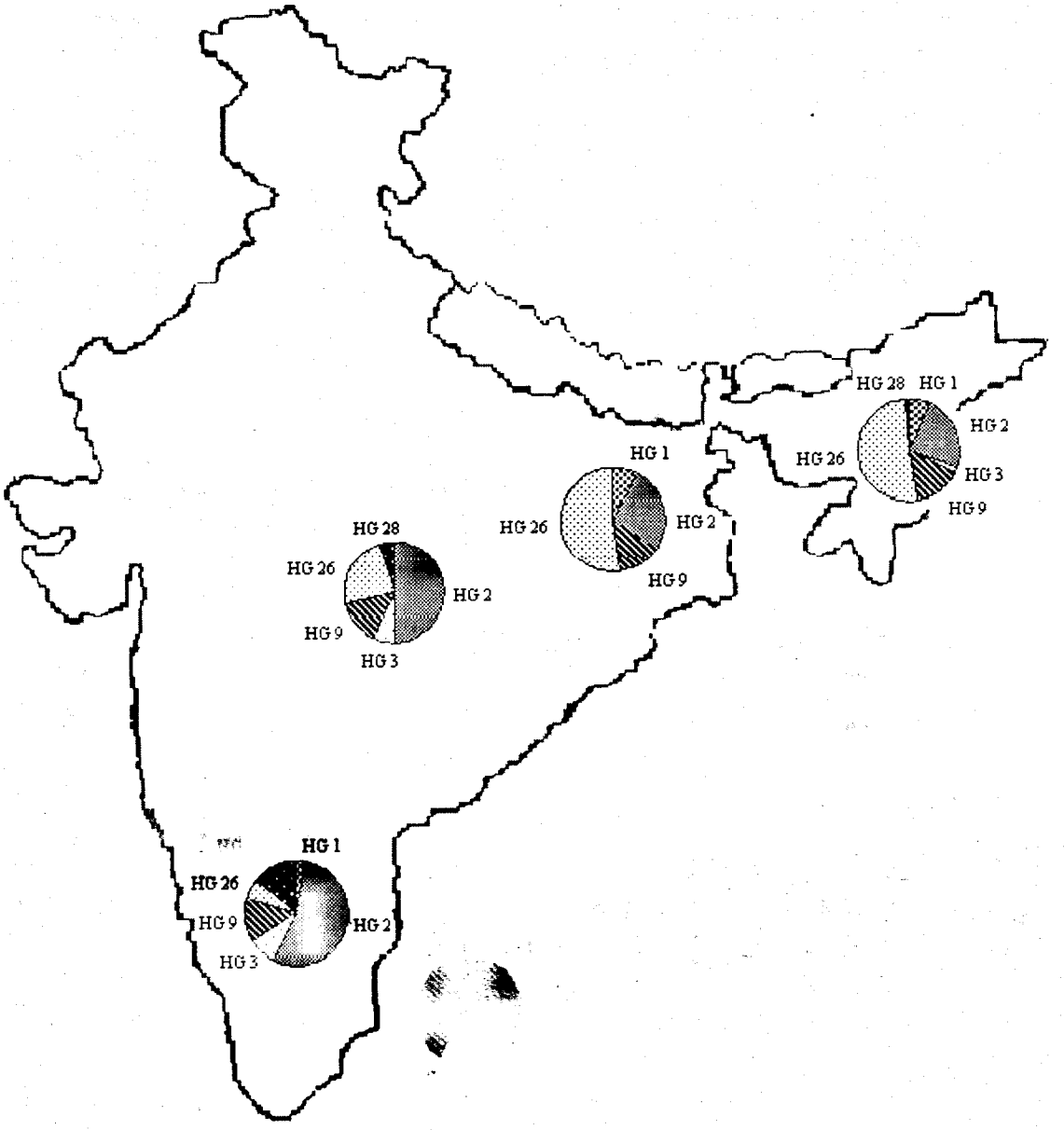


Figure 2.11(b)

(b)

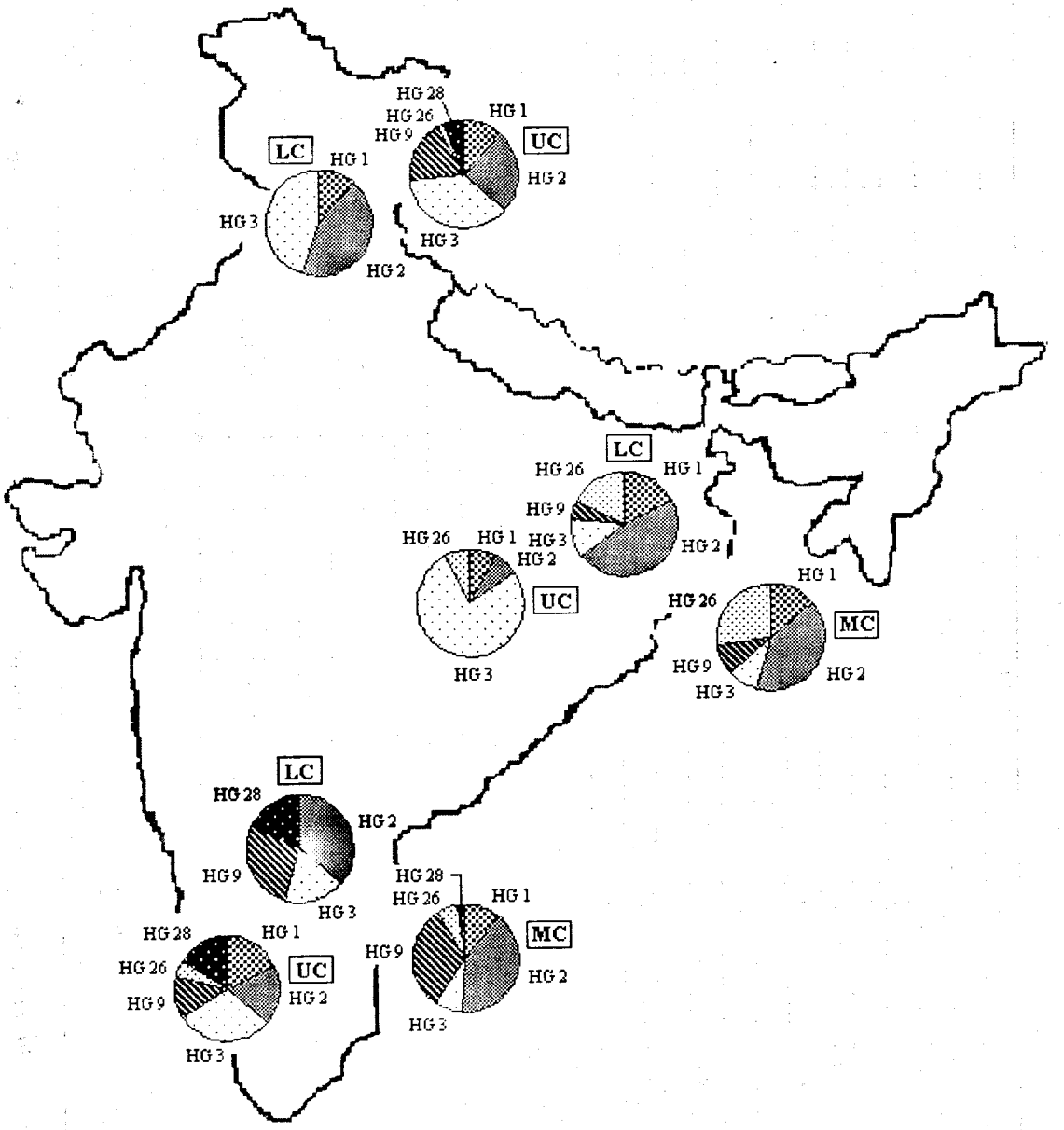


**Table 2.10**
**Percentage Frequencies of Alleles at STRPs on the Y-chromosome in Tribes and Castes Cross-classified by Language and Social Rank**

| LOCUS | REPEAT NO. | TRIBE | | | | CASTE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Austro-Asiatic | Dravidian | Tibeto-Burman | Indo-European | Dravidian | | | Indo-European | | |
| | | | | | | Upper | Middle | Lower | Upper | Middle | Lower |
| DYS19 | -13 | | | | | 2.5 | | | 1.4 | | |
| | 14 | | 1.1 | | | | 1.7 | | 15.5 | 15.4 | 2.8 |
| | 15 | 17.5 | 19.8 | 69.6 | | 40.0 | 24.1 | 25.0 | 39.4 | 26.9 | 48.6 |
| | 16 | 77.2 | 63.7 | 15.9 | 94.7 | 42.5 | 62.1 | 55.0 | 39.4 | 46.2 | 20.0 |
| | 17 | 3.5 | 13.2 | 7.2 | 5.3 | 12.5 | 5.2 | 15.0 | 4.2 | 11.5 | 28.6 |
| | 18 | 1.8 | 2.2 | 7.2 | | 2.5 | 6.9 | 5.0 | | | |
| | *n* | 57 | 91 | 69 | 19 | 40 | 58 | 20 | 71 | 26 | 35 |
| DYS388 | 8 | | | | | 2.5 | 1.7 | | | | |
| | 9 | | | 2.9 | | | | | | | 17.1 |
| | 10 | | | 52.2 | | | 1.7 | | 1.4 | | |
| | 11 | | 1.1 | 7.2 | | | | 5.0 | | | 1.8 |
| | 12 | 87.7 | 79.1 | 26.1 | 65.0 | 77.5 | 65.5 | 70.0 | 71.8 | 61.5 | 65.7 |
| | 13 | 1.8 | 7.7 | 11.6 | 30.0 | 5.0 | 6.9 | 10.0 | 12.7 | 26.9 | 14.3 |
| | 14 | | 8.8 | | 5.0 | 7.5 | 3.4 | | 4.2 | 3.8 | |
| | 15 | 10.5 | 3.3 | | | 5.0 | 20.7 | 15.0 | 9.8 | 3.8 | 2.8 |
| | 16 | | | | | 2.5 | | | | 3.8 | |
| | *n* | 57 | 91 | 69 | 20 | 40 | 58 | 20 | 71 | 26 | 35 |
| DYS389I | 7 | | 1.1 | | | 5.1 | 1.7 | | | | |
| | 8 | | | 3.7 | | | | | | | |
| | 9 | 14.0 | 25.3 | 51.8 | 26.3 | 28.2 | 36.2 | 40.0 | | 7.0 | 47.0 |
| | 10 | 68.4 | 51.6 | 19.8 | 21.0 | 35.9 | 39.7 | 35.0 | 33.3 | 69.2 | 23.5 |
| | 11 | 14.0 | 19.8 | 9.9 | 52.6 | 23.1 | 19.0 | 25.0 | 66.7 | 23.1 | 29.4 |
| | 12 | | 1.1 | 1.2 | | 5.1 | 1.7 | | | | |
| | 13 | | 1.1 | 3.7 | | 2.6 | | | | | |
| | 14 | 3.6 | | 9.9 | | | 1.7 | | | | |
| | *n* | 57 | 91 | 81 | 19 | 39 | 58 | 20 | 18 | 26 | 17 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DYS389II | 22 | | | 1.4 | | | | | | | |
| | 24 | | | 2.7 | | | | | | | |
| | 25 | | 5.5 | 12.2 | | 2.6 | 5.2 | | | 3.8 | 35.3 |
| | 26 | 8.8 | 23.1 | 50.0 | 15.0 | 23.1 | 34.5 | 25.0 | 11.1 | 3.8 | 5.9 |
| | 27 | 57.9 | 33.0 | 18.9 | 20.0 | 25.6 | 15.5 | 30.0 | 5.6 | 53.8 | 17.6 |
| | 28 | 22.8 | 33.0 | 8.1 | 35.0 | 20.5 | 19.0 | 25.0 | 11.1 | 23.1 | 23.5 |
| | 29 | 7.0 | 4.4 | 4.0 | 20.0 | 17.9 | 12.1 | 15.0 | 50.0 | 15.4 | 17.6 |
| | 30 | | 1.1 | 1.4 | 10.0 | 5.1 | 10.3 | | 16.7 | | |
| | 31 | 1.8 | | 1.4 | | 2.6 | 1.7 | | | | |
| | 32 | 1.8 | | | | 2.6 | 1.7 | 5.0 | 5.6 | | |
| | *n* | 57 | 91 | 74 | 20 | 39 | 58 | 20 | 18 | 26 | 17 |
| DYS390 | 21 | | 9.9 | | 20.0 | 2.5 | 1.7 | 5.0 | 2.9 | 7.7 | 2.9 |
| | 22 | 15.8 | 57.1 | 14.3 | 25.0 | 37.5 | 37.9 | 40.0 | 20.3 | 42.3 | 25.7 |
| | 23 | 26.3 | 20.9 | 18.2 | 5.0 | 25.0 | 29.3 | 15.0 | 18.8 | 26.9 | 31.4 |
| | 24 | 19.3 | 4.4 | 59.7 | 25.0 | 15.0 | 18.9 | 25.0 | 17.4 | 11.5 | 20.0 |
| | 25 | 38.6 | 7.7 | 7.8 | 25.0 | 17.5 | 12.1 | 15.0 | 39.1 | 11.5 | 20.0 |
| | 26 | | | | | 2.5 | | | 1.4 | | |
| | *n* | 57 | 91 | 77 | 20 | 40 | 58 | 20 | 69 | 26 | 35 |
| DYS391 | 8 | | | | | | | | | 3.8 | |
| | 9 | | 4.4 | | | | | 10.0 | | | |
| | 10 | 77.2 | 89.0 | 91.0 | 75.0 | 80.0 | 82.8 | 80.0 | 64.3 | 84.6 | 71.4 |
| | 11 | 19.3 | 6.6 | 9.0 | 25.0 | 17.5 | 17.2 | 10.0 | 31.4 | 11.5 | 28.6 |
| | 12 | 3.5 | | | | 2.5 | | | 4.3 | | |
| | *n* | 57 | 91 | 78 | 20 | 40 | 58 | 20 | 70 | 26 | 35 |
| DYS392 | 10 | | 9.7 | | | 7.9 | 5.6 | | 7.0 | | 5.6 |
| | 11 | 22.2 | 82.2 | 91.7 | 68.8 | 73.7 | 66.7 | 70.6 | 59.2 | 100.0 | 77.8 |
| | 12 | 2.8 | 3.2 | | 12.5 | 2.6 | 11.1 | | 1.4 | | 16.7 |
| | 13 | 16.7 | | | 18.7 | 2.6 | | 5.9 | 1.4 | | |
| | 14 | 58.3 | 4.8 | 8.3 | | 13.2 | 16.7 | 17.6 | 31.0 | | |
| | 15 | | | | | | | 5.9 | | | |
| | *n* | 36 | 62 | 12 | 16 | 38 | 18 | 17 | 71 | 1 | 18 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DYS393 | 8 | | | 1.2 | | | | | | | |
| | 10 | | | 4.8 | | 2.5 | 1.8 | | 0.0 | | |
| | 11 | | 8.8 | 4.8 | | 17.5 | 12.3 | 30.0 | 1.4 | | |
| | 12 | 17.8 | 28.6 | 61.4 | 55.0 | 22.5 | 33.3 | 30.0 | 32.4 | 57.7 | 25.7 |
| | 13 | 8.9 | 40.6 | 6.0 | 15.0 | 42.5 | 49.1 | 35.0 | 33.8 | 11.6 | 48.6 |
| | 14 | 25.0 | 17.6 | 16.9 | 20.0 | 5.0 | 3.5 | 5.0 | 5.6 | 30.8 | 25.7 |
| | 15 | 48.2 | 3.3 | 2.4 | 10.0 | 10.0 | | | 26.8 | | |
| | 17 | | | 2.4 | | | | | | | |
| | 18 | | 1.1 | | | | | | | | |
| | *n* | 56 | 91 | 83 | 20 | 40 | 57 | 20 | 71 | 26 | 35 |
| DYS425 | 5 | | | | | | 1.7 | | | | |
| | 6 | | | 1.2 | | | | | | | |
| | 7 | 3.5 | | 1.2 | | | | | 5.6 | | |
| | 8 | 15.8 | 13.2 | 2.4 | 10.0 | 2.5 | 5.2 | 5.0 | | 3.8 | |
| | 9 | 35.1 | 2.2 | 11.9 | 10.0 | | 1.7 | | | | 5.9 |
| | 10 | 36.8 | 29.7 | 71.4 | 35.0 | 60.0 | 67.2 | 55.0 | 88.9 | 76.9 | 82.3 |
| | 11 | 8.8 | 54.9 | 6.0 | 45.0 | 32.5 | 20.7 | 40.0 | 5.6 | 19.2 | 11.8 |
| | 13 | | | 6.0 | | 5.0 | 3.4 | | | | |
| | *n* | 57 | 91 | 84 | 20 | 40 | 58 | 20 | 18 | 26 | 17 |
| DYS426 | 8 | | 2.2 | 6.7 | | | 1.7 | | | | |
| | 9 | 86.0 | 85.7 | 83.1 | 70.0 | 55.0 | 81.0 | 70.0 | 5.9 | 57.7 | 64.7 |
| | 10 | 12.3 | 11.0 | 7.9 | 30.0 | 45.0 | 17.2 | 30.0 | 94.1 | 38.5 | 35.3 |
| | 11 | | 1.1 | | | | | | | 3.8 | |
| | 12 | | | 1.1 | | | | | | | |
| | 14 | 1.8 | | | | | | | | | |
| | 15 | | | 1.1 | | | | | | | |
| | *n* | 57 | 91 | 89 | 20 | 40 | 58 | 20 | 17 | 26 | 17 |

among whom the 9-repeats allele is the modal one. The number of alleles observed in the Indian populations (8 alleles) at this locus is also quite large compared to other global populations. These findings are largely similar for the DYS389II locus. At the DYS390 locus, the Dravidian and the Tibeto-Burman tribals stand out in that both these subgroups have clear modal alleles, containing 22- and 24-repeats respectively, while populations belonging to the other subgroups do not possess sharp modes. The 22-repeats allele is quite frequent in West Asia (Nebel et al. 2000), while in Central Asia the modal allele is the 25-repeats allele (Perez-Lezaun et al. 1999). All subgroups of Indian populations have the same modal allele at the DYS391 and DYS426 loci. At the DYS392 locus, the Austro-Asiatic tribals stand out. This subgroup has the 14-repeats allele as the modal one, while populations belonging to all other subgroups have the 11-repeats allele as the modal allele. The 14-repeats allele is infrequent in both West and Central Asia. At the DYS393 locus, allele containing 12 and 13 repeats are the modal ones in West and Central Asian populations, respectively. While these alleles have high, but variable, frequencies across subgroups of Indian populations, the Austro-Asiatic tribals again stand out in that they possess the 15-repeats allele at the highest frequency. At the DYS425 locus, Indian population subgroups have 10 or 11-repeats allele as the modal allele, except for the Austro-Asiatics tribals who possess the 9 and 10-repeats alleles at nearly equal frequencies (35%). Our findings at most of these STRP loci are consistent with data from populations of Pakistan (Mohyuddin et al. 2001). There are two striking exceptions. While among Pakistani populations the 12-repeats allele is the most predominant allele at the DYS425 locus, this allele was not observed in Indian populations. Further, at the DYS426 locus, the modal allele in India is the one containing 9 repeats, while in Pakistan this allele is virtually absent and the alleles containing 11 and 12 repeats are almost equally frequent.

To test whether the frequency distributions of repeat numbers were the same across relevant subsets of the cross-classified subgroups considered in Table 2.10, we have carried out Kruskal-Wallis tests. We have compared the distributions among (a) tribals belonging to the three linguistic groups (Austro-Asiatic, Dravidian and Tibeto-Burman), (b) upper middle and lower castes within each linguistic group (Dravidian and Indo-European), and (c) Dravidian and Indo-European castes belonging to the three ranks (upper middle and lower). All comparisons turned out to be statistically significant ($p<.005$), showing that there are significant differences in

the distributions of frequencies of repeat numbers at the various STRP loci among the various subsets of populations.

Table 2.11 provides the frequency distributions of 10-locus STRP haplotypes among tribals and caste populations, classified by linguistic affinity. Because data on all STRP loci were not available on all individuals, the sample sizes in the different cross-classified subgroups are small. A total of 143 haplotypes were observed in the set of 194 individuals on whom complete data on all 10 loci were available. Only 3 (2%) haplotypes are shared between tribals and castes. It is seen from this table that each subgroup of individuals possesses different sets of haplotypes; in other words, the sets of haplotypes in the different subgroups are virtually disjoint. (This feature was also seen in respect of 5-locus haplotypes, comprising the loci DYS19, DYS388, DYS390, DYS391 and DYS393, for which data were available on practically all (n=451) individuals screened for the Y-chromosomal loci. Data, not shown, are available on request.) We have also searched for haplotype sharing among various other groupings of the populations, including castes and tribes cross-classified by geographical region of habitat, castes within geographical zones classified by social rank, etc., but have consistently noted that there is virtually no sharing of haplotypes.

We have also examined haplotype sharing among Y-HGs. For the 10-locus data, of the 143 haplotypes observed in 194 individuals, only 5 (3.5%) haplotypes are shared across haplogroups. Thus, even across Y-HGs there is virtually no sharing of STRP haplotypes (data not shown; available upon request). The most striking features of the frequency distributions of STRP haplotypes among Y-HGs are: (i) the two haplotypes that are different by more than one repeat number at 4 loci, 16-12-10-26-25-10-14-15-9-9 and 16-12-11-27-22-10-11-13-11-9, are shared by 15 (20%) and 6 (8%) individuals within HG-2 ($n$=76); these haplotypes are not shared by individuals belonging to other HGs, (ii) one haplotype, 16-12-10-26-25-10-14-15-9-9, is shared by 7 (17%) individuals belonging to HG-26 ($n$=42); this haplotype is also not shared by individuals belonging to other HGs. (When we considered the 5-locus haplotype data, a total of 167 haplotypes were observed among 451 individuals, of which only 44 (26.3%) haplotypes were shared across haplogroups.) Interestingly, the two frequent haplotypes (16-12-10-26-25-10-14-15-9-9 and 16-12-11-27-22-10-11-13-11-9) within HG-2 are primarily concentrated within the Dravidian tribal group. Haplotype 16-12-10-26-25-10-14-15-9-9 is shared by 8

Table 2.11

**Distribution of STRP Haplotypes in Tribal and Caste Populations of India, Belonging to Different Linguistic Backgrounds**

| HAPLOTYPE* | TRIBE | | | | CASTE | |
|---|---|---|---|---|---|---|
| | Austro-Asiatic | Dravidian | Tibeto-Burman | Indo-European | Dravidian | Indo-European |
| 14-12- 7-24-23-10-10-13-10-10 | | | | | 1 | |
| 15-10- 9-24-24-10-11-12-10- 9 | | | 1 | | | |
| 15-10- 9-25-23-10-11-12-10- 9 | | | 1 | | | |
| 15-10- 9-25-23-10-14-11-10- 9 | | | 1 | | | |
| 15-10- 9-25-24-10-11-12-10- 9 | | | 2 | | | |
| 15-12- 7-25-23- 9-10-14-10-10 | | 1 | | | | |
| 15-12- 7-26-23-10-10-13-10-10 | | | | | | |
| 15-12- 9-24-22-10-11-11-11- 9 | | 1 | | | | |
| 15-12- 9-25-22-10-11-11-11- 9 | | 1 | | | 2 | |
| 15-12- 9-25-22-10-14-10-11- 9 | | | | | 1 | |
| 15-12- 9-25-22-10-14-11-11- 9 | | 2 | | | 4 | |
| 15-12- 9-25-22-10-14-13-11- 9 | | 1 | | | | |
| 15-12- 9-25-22-10-15-11-11- 9 | | | | | 1 | |
| 15-12- 9-26-22-10-11-11-11- 9 | | 1 | | | | |
| 15-12- 9-26-23-10-11-14-10- 9 | 1 | | | | | |
| 15-12- 9-26-23-11-11-13-10- 9 | 1 | | | | | |
| 15-12- 9-26-24-10-11-14-10- 9 | 1 | | | | | |
| 15-12-10-25-23- 9-10-18-10-10 | | 1 | | | | |
| 15-12-10-26-22-10-11-12-11- 9 | | | | | 1 | |
| 15-12-10-26-23-10-11-15-10-10 | | | | | 1 | |
| 15-12-10-27-22-10-11-13-10- 9 | | | | | 1 | |
| 15-12-10-27-23-10-11-14- 8- 9 | | 1 | | | | |
| 15-12-10-27-23-11-12-15-10-10 | 1 | | | | | |
| 15-12-10-28-24-11-12-13-10- 9 | | | | | 1 | |
| 15-12-10-28-25-10-14-15-10-10 | | | | | | 1 |
| 15-12-10-29-24-10-11-13-10- 9 | | | | | 1 | |
| 15-12-10-29-24-11-11-13-10- 9 | | | | | 2 | |
| 15-12-11-27-24-10-13-12- 8-10 | | | | | 1 | |
| 15-12-14-30-23-10-11-14- 7- 9 | 1 | | | | | |
| 15-13-10-25-23-10-10-14-10-10 | | 1 | | | | |
| 15-14-10-27-23-10-11-11-10- 9 | | | | | 1 | |
| 15-14-10-27-23-10-11-12-11- 9 | | | | | 1 | |
| 15-14-13-26-23-10-11-12-11- 9 | | | | | 1 | |
| 15-15-11-27-23-10-11-12-10- 9 | | | | | 1 | |
| 15-15-11-27-23-10-11-13-10- 9 | | 1 | | | | |
| 15-15-11-27-23-11-11-15-10- 9 | | 1 | | | | |
| 15-16-10-27-22-10-11-13-10- 9 | | | | | 1 | |
| 16- 8-11-28-21-11-11-12-11- 9 | | | | | 1 | |
| 16-10- 9-24-24-10-11-12-10- 9 | | | 1 | | | |
| 16-11-10-27-24-10-11-13-10- 9 | | 1 | | | | |
| 16-11-11-31-22-10-11-13- 8- 9 | | | | | 1 | |
| 16-12- 9-24-22-10-11-15-11- 9 | | 1 | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16-12- 9-25-22-10-11-11-11- 9 | | 1 | | | | |
| 16-12- 9-25-22-10-11-12-11- 9 | | | | 3 | | |
| 16-12- 9-25-22-10-11-15-13- 9 | | | | | 1 | |
| 16-12- 9-25-22-10-14-11-11- 9 | | | | | 1 | |
| 16-12- 9-25-25-10-11-13- 8- 9 | | 2 | | | | |
| 16-12- 9-26-22-10-14-15-11- 9 | 1 | | | | | |
| 16-12- 9-26-23-10-10-15-10-10 | | | | | 1 | |
| 16-12- 9-26-24-12-11-12-10- 9 | | | | | 1 | |
| 16-12- 9-27-22-10-14-11-11- 9 | | | | | 1 | |
| 16-12- 9-27-24-10-11-13-13- 9 | | | | | 1 | |
| 16-12- 9-27-25-10-14-15- 9- 9 | 1 | | | | | |
| 16-12-10-24-22-10-11-12- 8- 9 | | 1 | | | | |
| 16-12-10-25-24-10-13-15- 9- 9 | 1 | | | | | |
| 16-12-10-26-21-10-11-12-11- 9 | | | | | 1 | |
| 16-12-10-26-22-10-11-12-11- 9 | | 14 | | | 1 | |
| 16-12-10-26-22-10-11-13-11- 9 | 1 | | | | | |
| 16-12-10-26-22-10-14-12-11- 9 | | | | | 1 | |
| 16-12-10-26-22-10-14-15- 9- 9 | 1 | | | | | |
| 16-12-10-26-23-10-10-13-10-10 | | 1 | | | | |
| 16-12-10-26-23-10-11-13-10-10 | | | | | 1 | |
| 16-12-10-26-23-10-14-15-10-10 | | | | | | 1 |
| 16-12-10-26-25-10-13-14- 9- 9 | 2 | | | | | |
| 16-12-10-26-25-10-13-15- 9- 9 | | | | 1 | | |
| 16-12-10-26-25-10-14-13- 9- 9 | 1 | | | | | |
| 16-12-10-26-25-10-14-15- 9- 9 | 7 | | | | | |
| 16-12-10-27-22-10-11-12-11- 9 | | 1 | | 1 | | |
| 16-12-10-27-22-10-11-13-11- 9 | | 1 | | | | |
| 16-12-10-27-25-10-11-13-10-10 | | | | | 1 | |
| 16-12-10-27-25-10-11-14- 8- 9 | 1 | | | | | |
| 16-12-10-27-25-10-13-14- 8- 9 | 1 | | | | | |
| 16-12-10-27-25-11-13-14- 8- 9 | | | | 2 | | |
| 16-12-10-27-25-11-14-15- 8- 9 | 2 | | | | | |
| 16-12-10-28-24-10-14-15- 8- 9 | 1 | | | | | |
| 16-12-10-28-25-11-14-15-10-10 | | | | | | 1 |
| 16-12-10-29-23-10-11-14-10-10 | | 1 | | | | |
| 16-12-10-29-25-10-14-15-10-10 | | | | | | 1 |
| 16-12-10-31-22-10-14-11- 5- 9 | | | | | 1 | |
| 16-12-11-26-22-11-14-15-10- 9 | 1 | | | | | |
| 16-12-11-27-21-10-11-13-11- 9 | | | | 1 | | |
| 16-12-11-27-21-11-12-13-11- 9 | | | | 1 | | |
| 16-12-11-27-22-10-11-12-11- 9 | | | | | 1 | |
| 16-12-11-27-22-10-11-13-11- 9 | | 6 | | | | |
| 16-12-11-27-22-10-14-15-11- 9 | | | | | | 1 |
| 16-12-11-27-23-10-11-12-11- 9 | | | | 1 | | |
| 16-12-11-27-23-10-14-15-10-10 | | | | | | 1 |
| 16-12-11-28-22-10-11-12-11- 9 | 1 | | | | | |
| 16-12-11-28-23-10-11-13-10-10 | | | | | 1 | |
| 16-12-11-28-24-10-11-13-10-10 | | | | | 3 | |
| 16-12-11-28-24-10-11-14-10-10 | | | | | 1 | |
| 16-12-11-28-24-11-14-15- 8- 9 | 2 | | | | | |
| 16-12-11-28-25-10-14-15-10-10 | | | | | | 1 |
| 16-12-11-28-25-11-11-13-10-10 | | | | | 2 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16-12-11-28-25-11-14-15-10-10 | | | | | | 2 |
| 16-12-11-28-25-12-14-15-10-10 | | | | | | 1 |
| 16-12-11-29-24-10-11-13-10-10 | | | | | 1 | |
| 16-12-11-29-25-10-14-15-10-10 | | | | | | 1 |
| 16-12-11-31-25-10-11-13-10-10 | | | | | 1 | |
| 16-12-12-29-25-10-11-13-10-10 | | | | | 1 | |
| 16-12-13-25-22-10-11-12- 9- 8 | | 1 | | | | |
| 16-12-14-31-25-10-14-15- 7-14 | 1 | | | | | |
| 16-13- 7-27-23-10-12-14-10- 9 | | | | | 1 | |
| 16-13-10-25-23-10-14-15-10-10 | | | | | | 2 |
| 16-13-10-26-24-10-11-13-10- 9 | | | | | 1 | |
| 16-13-11-26-22-10-11-13-10- 9 | | | | | 1 | |
| 16-13-11-26-24-10-11-12-10-10 | | | | 1 | | |
| 16-13-11-27-21-11-12-13-10- 9 | | | | | 1 | |
| 16-13-11-28-24-10-11-12-10-10 | | | | 2 | | |
| 16-13-11-29-25-10-11-12-10-10 | | | | 1 | | |
| 16-14- 9-25-21-10-11-13-10- 9 | | 1 | | | | |
| 16-14- 9-25-23-11-12-13-10- 9 | | 1 | | | | |
| 16-14- 9-25-24-10-12-13-10- 9 | | 1 | | | | |
| 16-14- 9-26-21-10-11-13-10- 9 | | 2 | | | | |
| 16-14-10-24-22- 9-11-15-10-11 | | 1 | | | | |
| 16-14-11-28-21-11-11-13-11- 9 | | | | 1 | | |
| 16-15- 9-24-23-10-11-12-10- 9 | | | | | 1 | |
| 16-15- 9-24-24-10-11-12-10- 9 | | | | | 1 | |
| 16-15- 9-25-23-10-11-11-10- 9 | | | | | 1 | |
| 16-15- 9-25-24-10-11-12-10- 9 | | | | | 1 | |
| 16-15- 9-26-24- 9-11-12-10- 9 | | | | | 1 | |
| 16-15-10-26-23-10-11-12-10- 9 | 1 | | | | | |
| 16-15-10-26-23-10-11-13-10- 9 | | | | | 1 | |
| 17-12- 9-27-25-11-11-13-10-10 | | | | | 1 | |
| 17-12-10-26-23-10-11-13- 8- 9 | | 1 | | | | |
| 17-12-10-27-23-10-11-13- 8- 9 | | 2 | | | | |
| 17-12-10-27-23-10-11-13-10- 9 | | 1 | | | | |
| 17-12-10-27-23-10-11-13-11- 9 | | 1 | | | | |
| 17-12-10-27-24-11-11-13-10-10 | | | | | 1 | |
| 17-12-10-27-25-10-11-13-10-10 | | | | | 1 | |
| 17-12-10-27-25-10-14-14- 8- 9 | 1 | | | | | |
| 17-12-10-27-25-12-13-14- 8- 9 | 1 | | | | | |
| 17-12-10-28-25-10-11-13-10-10 | | | | | 1 | |
| 17-12-10-28-25-11-11-13-10-10 | | | | | 1 | |
| 17-12-11-28-25-10-11-13-10-10 | | 1 | | | | |
| 17-13-10-27-21-10-11-14- 8- 9 | | 2 | | | | |
| 17-13-10-27-21-10-11-14-10- 9 | | 1 | | | | |
| 17-13-10-30-25-10-11-14-10-10 | | | | | 1 | |
| 17-13-11-28-21-11-12-14-10- 9 | | | | 1 | | |
| 18-12- 9-27-22-10-11-14-13- 9 | | | | | 1 | |
| 18-12-10-27-25-10-13-15- 9- 9 | 1 | | | | | |
| 18-12-10-28-25-11-11-13-10-10 | | | | | 1 | |
| 18-12-11-28-26-10-11-13-10-10 | | | | | 1 | |
| Total | 34 | 58 | 6 | 16 | 63 | 13 |

Kamars and 6 Kotas, while haplotype 16-12-11-27-22-10-11-13-11-9 is shared by 6 Kurumbas. The most likely cause of such structured sharing of haplotypes is founder effect.

Based on the average variance of repeat numbers at the 10 STRP loci among individuals belonging to the same HG, we have estimated the ages and the 95% CIs of the HGs. These are provided in Table 2.12. Our age estimates are, by and large, in agreement earlier estimates (Rosser et al. 2000; Zerjal et al. 2002; Tajima et al. 2002).

HG-1 probably arose in Central Asia (Zerjal et al. 2002) and its age has been estimated to be between 19000 ybp (Tajima et al 2002) and 23000 ybp (Zerjal et al. 2002). We have estimated the age to be about 16000 ybp, with a 95% confidence interval of 9477-30000 ybp. We have estimated that HG-2 has an age similar to that of HG-1. However, HG-2 possibly comprises a heterogeneous set of loosely related chromosomes (Zerjal et al. 2002), and hence its age estimate may not carry much meaning. We have estimated that HG-3, which is found primarily in Central and Eastern Europe (Rosser et al. 2000), is a young haplogroup that probably arose about 5400 ybp. This finding agrees with those of Karafet et al. (1999) and Tajima et al. (2002). Similarly, our estimate of the age of HG-9 (about 15000 ybp) also agrees with that of Hammer et al. (2000). Our estimate of the age of HG-26 (20000 ybp) disagrees completely with that of Tajima et al.(2002), who have estimated its age to be $95000 \pm 51000$ ybp. The reason for this disagreement is unclear to us. However, our estimate of the age of HG-28 (about 14000 ybp) is consistent with the fact that HG-28 is a derivative of HG-26.

Phylogenetic analysis: In order to examine the nature and extent of similarities among the populations, we have constructed neighbor-joining trees based on frequencies of Y chromosomal haplogroups and STRPs. These are presented in Figures 2.12 and 2.13. The only notable feature that emerges from these figures is that the Tibeto-Burman tribal populations of the northeastern region, by and large, cluster themselves separately from the populations of other geographical regions or language groups. There is no clear clustering of populations either by social rank or by geographical region or by language affinity, except from the Tibeto-Burman speaking tribals.

In order to examine the nature of similarities of the Indian populations with those of Central and West Asia and Pakistan, we have collated data on Y-haplogroup frequencies from the published literature (Hammer et al. 2000; Nebel et al, 2000; Rosser et al. 2000l; Qamar et al. 2002) and have analyzed these jointly with the data generated in the present study. The neighbor-joining tree constructed on the basis of these data is presented in Figure 2.14. Several

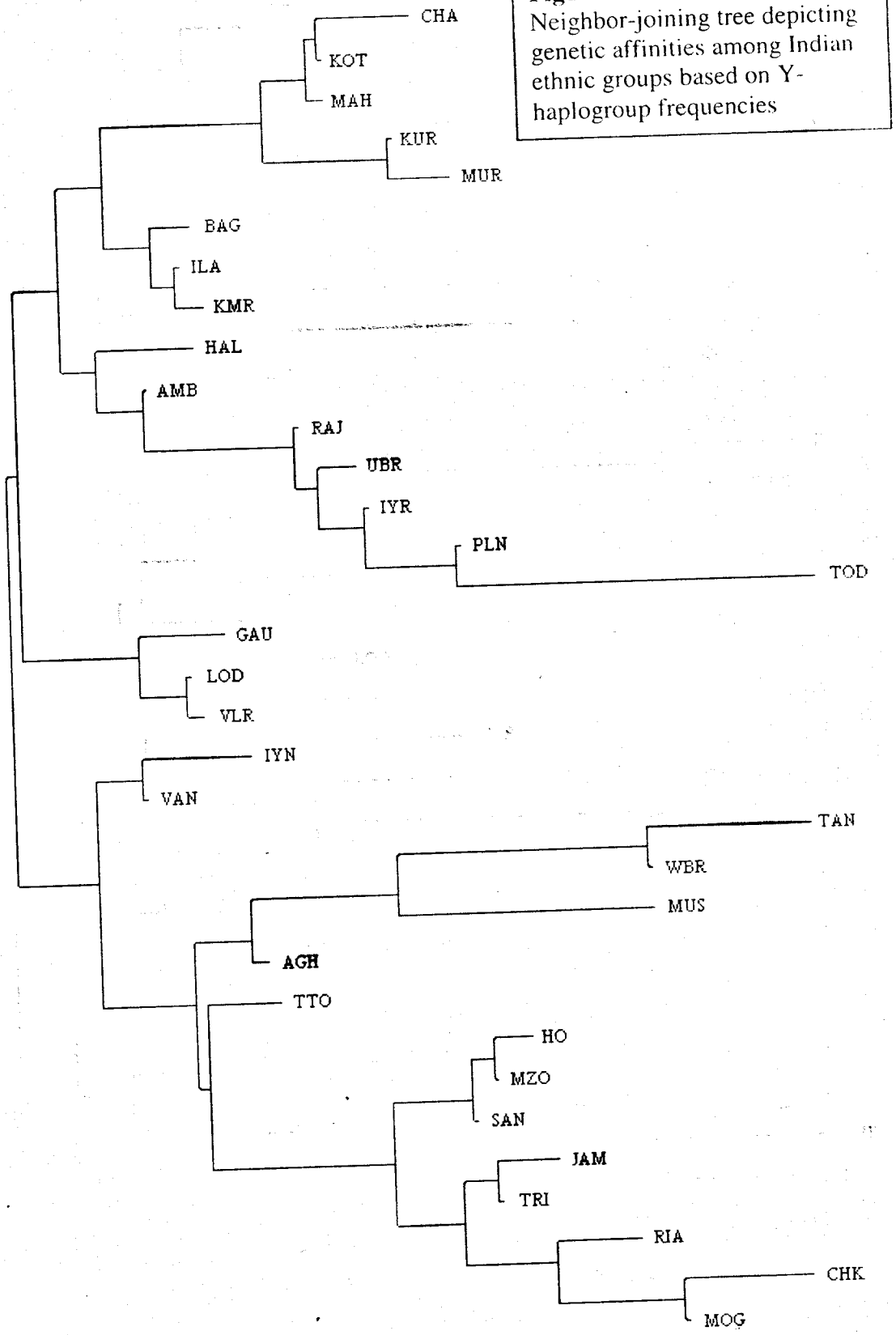



**Figure 2.12**
Neighbor-joining tree depicting genetic affinities among Indian ethnic groups based on Y-haplogroup frequencies

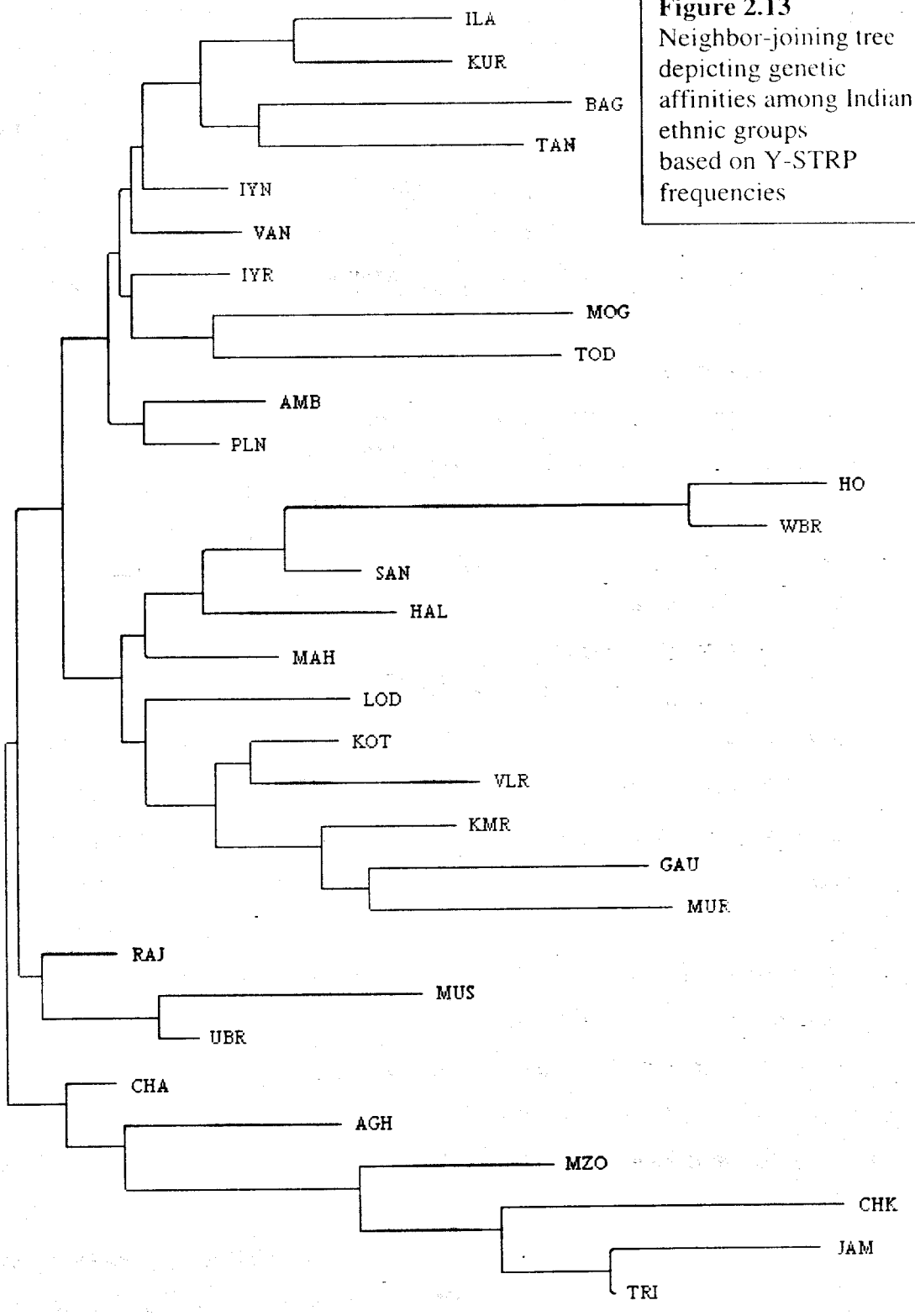



**Figure 2.13**
Neighbor-joining tree depicting genetic affinities among Indian ethnic groups based on Y-STRP frequencies

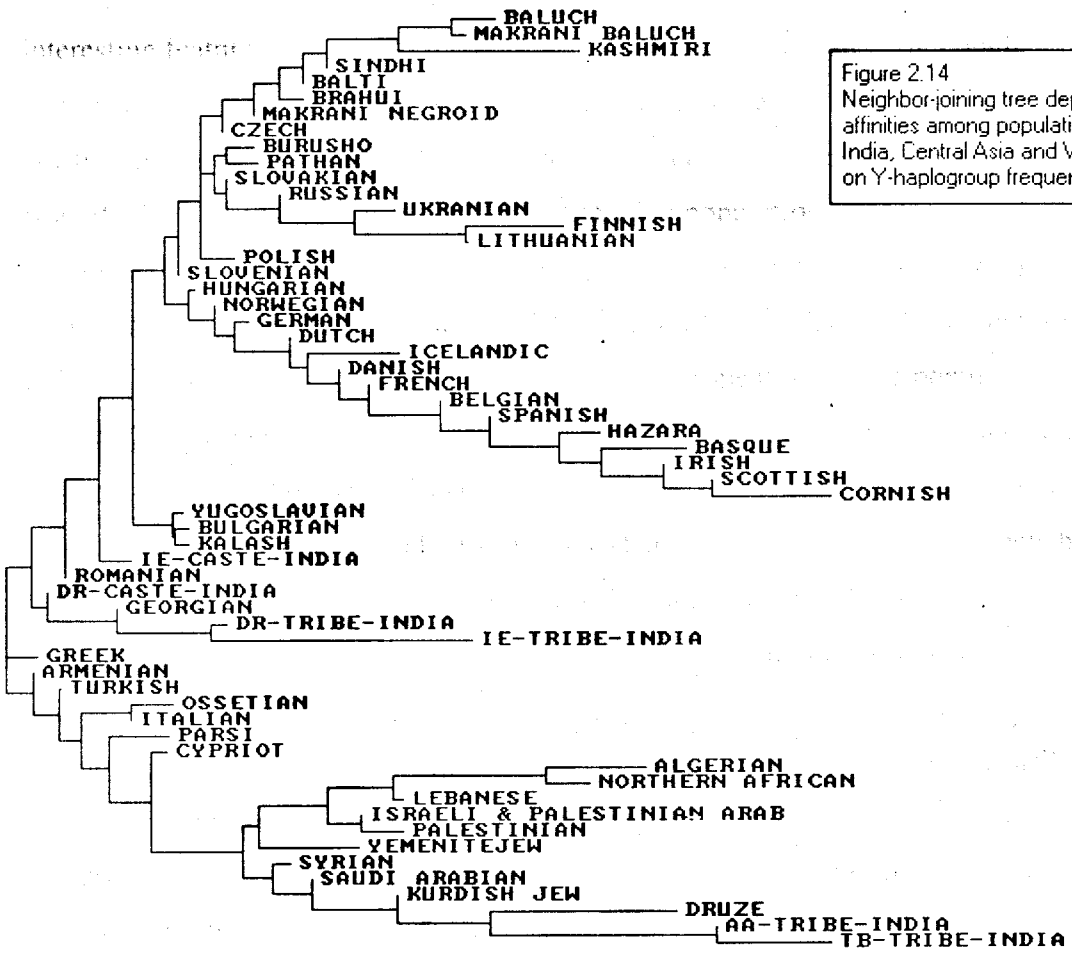


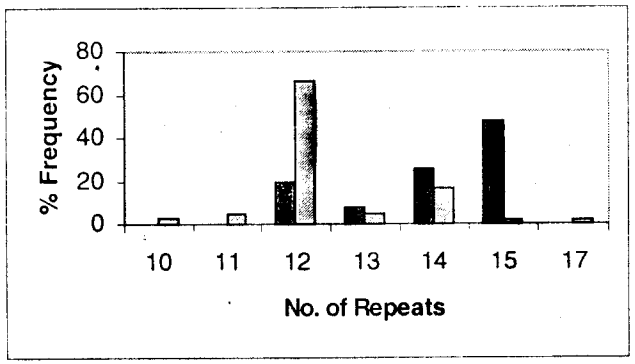

Figure 2.14
Neighbor-joining tree depicting genetic affinities among population groups of India, Central Asia and West Asia based on Y-haplogroup frequencies

interesting features emerge from this figure. First, the Dravidian caste populations are nearly equidistant from the Dravidian tribal and the Indo-European caste populations. Second, the Austro-Asiatic and the Tibeto-Burman tribal populations cluster together, and are very distant from the Dravidian or the Indo-European speaking populations of India. Third, the Austro-Asiatic and the Tibeto-Burman tribal populations belong to the cluster that also contain several West Asian and Northern African populations. Fourth, the Central Asian, but not West Asian, populations are closer to the Dravidian and Indo-European speaking populations of India. Fifth, although the Brahuis of Pakistan speak a dialect that belongs to the Dravidian language family which is often cited (Cavalli-Sforza 2000) as a major pointer to the possibility that the Dravidian language may have entered India from Iran (where, in the Elam region, dialects belonging to the Dravidian family are also spoken) through Pakistan, this population (Brahui) is very dissimilar from the Indian populations, and clusters with other populations of Pakistan and several European populations.

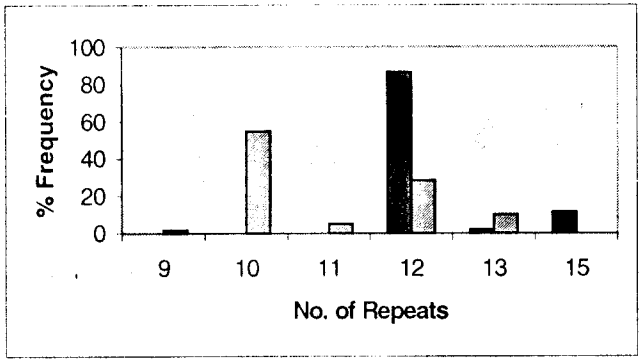
Austro-Asiatic and Tibeto-Burman Tribals: As mentioned earlier, there are striking similarities in the frequencies of Y-HGs between Austro-Asiatic and Tibeto-Burman tribals. HGs 3 and 26 are absent in both groups and the modal HG is 26 (50% and 70%, respectively, in the two groups). HG-26 occurs in low frequencies among tribals and castes of other linguistic affiliations. The haplogroup frequencies provide low discrimination between the Austro-Asiatic and Tibeto-Burman tribals. The probability of correctly classifying a tribal into one of these two groups is only 62.6%. As a matter of fact, the probability of correctly distinguishing an Austro-Asiatic tribal from a Tibeto-Burman tribal based on Y-haplogroups is only 50%. Therefore, a natural question is whether these tribals who now speak dialects that belong to different linguistic groups arose from a common stock of males or have had significant inflow of Y-chromosomes from a common population. To examine this issue, we have carried out a stepwise linear discriminant analysis using the 10 Y-STR markers. This analysis has revealed that the best discriminators between the Austro-Asiatic ($n$=52) and the Tibeto-Burman tribals ($n$=60) are, in decreasing order of importance, DYS393, DYS388 and DYS391. The overall probability of correct classification using these three STRPs is very high (79.5%). There is a higher percentage (28.3%) of Tibeto-Burman tribals misclassified as Austro-Asiatic tribals than of Austro-Asiatic tribals misclassified as Tibeto-Burman tribals (11.5%). To obtain a clearer idea, we have plotted the distributions of allele frequencies at these three discriminating loci (Figure 2.15). This figure

**Figure 2.15**
Frequency distributions alleles and haplotypes at three Y-chromosomal STR loci that discriminate between Austro-Asiatic (black) and Tibeto-Burman (grey) tribal populations of India

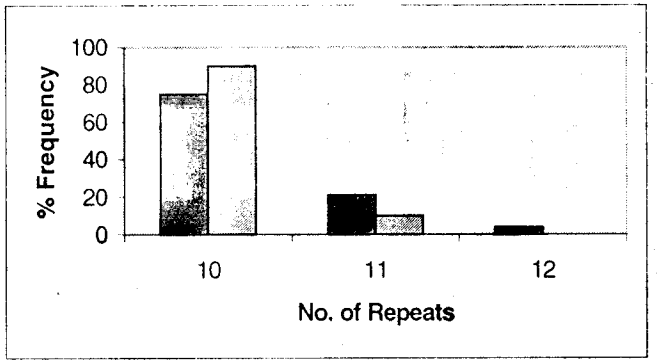
**DYS393**




**DYS388**




**DYS391**



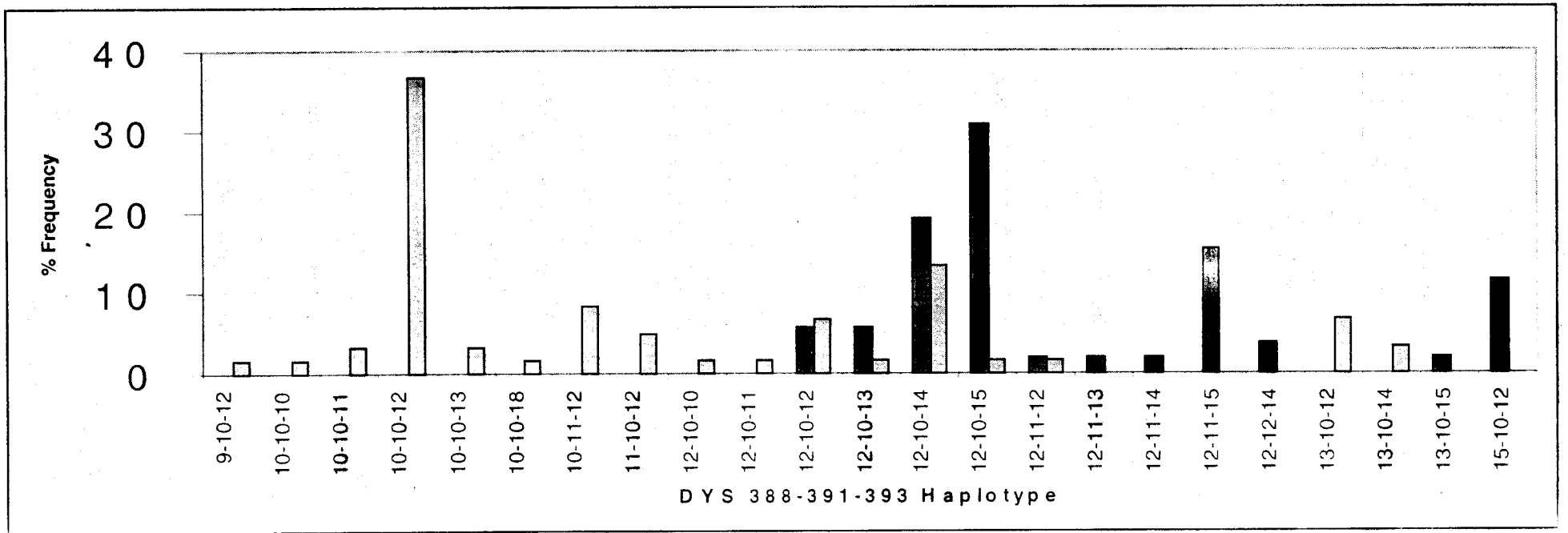



**Table 2.12**
**Estimated Ages of Various Y-chromosomal Haplogroups**

| ESTIMATE (YEARS) | HAPLOGROUP | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **9** | **26** | **28** |
| AGE | 16322 | 12508 | 5416 | 15300 | 20049 | 14290 |
| LOWER 95% CI | 9477 | 7263 | 3145 | 8884 | 11641 | 8297 |
| UPPER 95% CI | 29979 | 22974 | 9947 | 28103 | 36824 | 26246 |

shows that indeed there are dramatic differences between Austro-Asiatic and Tibeto-Burman tribals in respect of allele frequencies at these three loci. We have also computed the haplotype frequencies using data of these three loci among these two groups of tribals. These frequency distributions are also shown in Figure 2.15, which clearly reveals that the frequency distributions of haplotypes are substantially non-overlapping and the Kruskal-Wallis test reveals that they are significantly different ($p<.0001$). This indicates that the source populations from which the Austro-Asiatic and Tibeto-Burman tribals have been derived are not the same nor was there a large inflow of genes from a common population into both these groups.

Genetic Differentiation:

We have calculated the $F_{ST}$ values for various groupings of populations, separately in respect of Y-chromosomal RSP and STRP frequencies. These values are presented in Table 2.7. It is seen that with the exception of the upper caste groups, there is generally a good correpondance between the $F_{ST}$ values computed on the basis of the RSP and the STRP frequencies. The tribal populations belonging to the four linguistic groups or, geographical zones (which are confounded) are maximally differentiated. As genetic differentiation is expected to be higher among more ancient populations, this bolsters the view that the tribal populations are older than the caste populations. The degree of differentiation is quite substantial, especially in respect of Y-STRPs. There is a high rank correlation between the $F_{ST}$ values based on mt and Y polymorphic markers, although the extent of differentiation is substantially lower for mt markers for all groups of populations. Among upper castes across the geographical regions, the degree of differentiation with respect to Y-RSP markers is substantially lower than with respect to Y-STRP markers. Because STRP loci tend to evolve faster than RSP loci, this may be indicative of multiple waves of male admixture into the upper caste groups from the same populations or from populations derived from the same ancestral population.

## *Autosomal Polymorphisms*

Allele and haplotype frequencies: Population-wise estimates of allele frequencies at the various unlinked loci are given in Table 2.13. These estimates are generally within the ranges observed earlier in various Indian and other populations. Except for the *Alu*-D1 locus, there is no consistently significant deviation from Hardy-Weinberg proportions for any locus or for any

population. At the *Alu*-D1 locus, there is significant deviation of genotype frequencies from those expected under Hardy-Weinberg equilibrium in the vast majority of populations. The reason for this is unclear. While selection can cause such systematic deviations, selection is not known to operate at this locus. Population-wise estimates of haplotype frequencies at the three linked loci are provided in Table 2.14. While not all populations possess all the haplotypes, the Chakmas (CHK) stand out as distinct since at all the loci they possess only a small number of haplotypes. This is either because of founder effect or drift or both. Modal haplotype also vary across populations; this feature has been observed earlier (Kulozik et al. 1986; ALFRED database and Kidd et al. 1998). In fact, when the haplotype frequencies in the present populations are compared with the data presented in the ALFRED database, one finds some evidence of clinal gradients from West Asia into northern and central India. (These features will be presented in a separate publication.)

Phylogenetic analysis: We have examined the nature and extent of genetic similarities among the various populations based on estimated autosomal allele and haplotype frequencies presented in Tables 13 and 14. The neighbor-joining tree depicting the relationships among the study populations is given in Figure 2.16. Consistent with our findings on genetic affinities based on frequencies of Y haplogroups and STRPs, the autosomal data also reveal that the Tibeto-Burman speaking populations of the northeast form a distinct cluster. Interestingly, unlike other sets of markers (mt and Y), the data of autosomal markers reveal that the Austro-Asiatic tribal populations are also distinctive from the other populations. No clear clustering of the Dravidian and Indo-European speaking populations is discernible.

Population Structure Analysis: To explore in further detail the presence of cryptic population structure and the relationships among the various subgroups of populations, we have carried out a "structure" analysis, developed by Pritchard et al. (2000). In this analysis, a sample of individuals is assumed to have originated in one of K (unknown) populations. Based on the genotype data at one or several loci, each individual is assigned to a population, or jointly to two or more populations if the data are indicative of admixture, using a Bayesian technique of posterior probability computation. The number of populations, K, is also simultaneously estimated. This procedure provides an exploratory overview of the nature of genetic structuring of a population, or a set of populations. Using this procedure, we have analyzed data of (a)

Table 2.13
Sample Sizes, Estimated Allele Frequencies with Standard Errors (S.E.) and Hardy-Weinberg Chi-squared Values at 17 Unlinked Loci

| Popula-tion | Alu-ACE | | | | Alu-FXIIIB | | | | Alu-D1 | | | | Alu-APO | | | | Alu-CD4 | | | | Alu-PLAT | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ |
| AGH | 24 | 0.583 | 0.07 | 0.960 | 24 | 0.646 | 0.07 | 3.153 | 24 | 0.583 | 0.07 | 5.662 | 24 | 0.187 | 0.06 | 2.400 | 24 | 0.062 | 0.03 | 0.107 | 24 | 0.458 | 0.07 | 0.734 |
| AMB | 50 | 0.520 | 0.05 | 0.742 | 50 | 0.320 | 0.05 | 1.493 | 50 | 0.630 | 0.05 | 6.355 | 50 | 0.200 | 0.04 | 0.781 | 50 | 0.030 | 0.02 | 0.048 | 50 | 0.350 | 0.05 | 0.489 |
| BAG | 31 | 0.339 | 0.06 | 1.340 | 31 | 0.468 | 0.06 | 1.653 | 31 | 0.355 | 0.06 | 5.903 | 31 | 0.145 | 0.04 | 0.894 | 31 | 0.065 | 0.03 | 0.147 | 31 | 0.532 | 0.06 | 0.025 |
| CHA | 25 | 0.300 | 0.06 | 0.510 | 25 | 0.220 | 0.06 | 10.574 | 25 | 0.500 | 0.07 | 4.840 | 25 | 0.280 | 0.06 | 0.002 | 25 | 0.000 | 0.00 | 0.000 | 25 | 0.460 | 0.07 | 1.896 |
| CHK | 10 | 0.300 | 0.10 | 2.744 | 10 | 0.150 | 0.08 | 0.311 | 9 | 0.611 | 0.11 | 0.803 | 10 | 0.300 | 0.10 | 1.837 | 10 | 0.000 | 0.00 | 0.000 | 10 | 0.400 | 0.11 | 0.625 |
| GAU | 15 | 0.400 | 0.09 | 0.185 | 15 | 0.267 | 0.08 | 0.008 | 15 | 0.800 | 0.07 | 5.104 | 15 | 0.500 | 0.09 | 3.267 | 15 | 0.033 | 0.03 | 0.018 | 13 | 0.615 | 0.10 | 0.008 |
| HAL | 48 | 0.354 | 0.05 | 0.000 | 48 | 0.302 | 0.05 | 0.893 | 48 | 0.583 | 0.05 | 2.508 | 47 | 0.309 | 0.05 | 5.814 | 48 | 0.094 | 0.03 | 0.514 | 47 | 0.426 | 0.05 | 0.813 |
| HO | 54 | 0.269 | 0.04 | 0.588 | 53 | 0.198 | 0.04 | 0.633 | 52 | 0.548 | 0.05 | 17.072 | 53 | 0.151 | 0.03 | 1.675 | 54 | 0.000 | 0.00 | 0.000 | 54 | 0.287 | 0.04 | 0.928 |
| ILA | 50 | 0.250 | 0.04 | 0.009 | 50 | 0.360 | 0.05 | 0.870 | 50 | 0.400 | 0.05 | 22.222 | 50 | 0.430 | 0.05 | 1.678 | 50 | 0.040 | 0.02 | 0.087 | 50 | 0.450 | 0.05 | 0.005 |
| IYN | 50 | 0.460 | 0.05 | 2.158 | 51 | 0.373 | 0.05 | 0.305 | 51 | 0.539 | 0.05 | 16.336 | 51 | 0.333 | 0.05 | 5.338 | 51 | 0.108 | 0.03 | 0.745 | 51 | 0.441 | 0.05 | 0.372 |
| IYR | 50 | 0.430 | 0.05 | 3.506 | 50 | 0.450 | 0.05 | 0.250 | 50 | 0.580 | 0.05 | 17.375 | 50 | 0.140 | 0.03 | 1.435 | 50 | 0.110 | 0.03 | 0.764 | 50 | 0.490 | 0.05 | 0.317 |
| JAM | 52 | 0.317 | 0.05 | 0.023 | 55 | 0.245 | 0.04 | 0.250 | 55 | 0.591 | 0.05 | 18.908 | 55 | 0.173 | 0.04 | 2.398 | 55 | 0.018 | 0.01 | 0.019 | 55 | 0.555 | 0.05 | 1.091 |
| KMR | 57 | 0.360 | 0.04 | 0.876 | 57 | 0.254 | 0.04 | 1.390 | 57 | 0.658 | 0.04 | 9.835 | 57 | 0.351 | 0.04 | 3.008 | 57 | 0.018 | 0.01 | 0.018 | 57 | 0.465 | 0.05 | 0.029 |
| KOT | 45 | 0.378 | 0.05 | 2.360 | 45 | 0.122 | 0.03 | 0.207 | 45 | 0.411 | 0.05 | 33.461 | 45 | 0.233 | 0.04 | 0.141 | 45 | 0.000 | 0.00 | 0.000 | 44 | 0.341 | 0.05 | 0.006 |
| KUR | 54 | 0.194 | 0.04 | 0.001 | 54 | 0.306 | 0.04 | 0.001 | 54 | 0.472 | 0.05 | 29.565 | 54 | 0.417 | 0.05 | 0.122 | 54 | 0.000 | 0.00 | 0.000 | 54 | 0.296 | 0.04 | 1.291 |
| LOD | 32 | 0.141 | 0.04 | 0.857 | 32 | 0.172 | 0.05 | 0.005 | 32 | 0.719 | 0.06 | 9.201 | 32 | 0.547 | 0.06 | 0.166 | 32 | 0.016 | 0.02 | 0.008 | 32 | 0.375 | 0.06 | 1.280 |
| MAH | 34 | 0.441 | 0.06 | 0.185 | 34 | 0.338 | 0.06 | 2.614 | 34 | 0.412 | 0.06 | 8.993 | 34 | 0.176 | 0.05 | 1.561 | 34 | 0.162 | 0.04 | 1.972 | 34 | 0.500 | 0.06 | 1.059 |
| MOG | 25 | 0.420 | 0.07 | 0.113 | 25 | 0.260 | 0.06 | 0.104 | 24 | 0.521 | 0.07 | 13.484 | 25 | 0.220 | 0.06 | 0.848 | 25 | 0.000 | 0.00 | 0.000 | 24 | 0.375 | 0.07 | 0.107 |
| MUR | 49 | 0.469 | 0.05 | 7.567 | 49 | 0.214 | 0.04 | 2.205 | 49 | 0.653 | 0.05 | 6.689 | 49 | 0.286 | 0.05 | 7.840 | 49 | 0.010 | 0.01 | 0.005 | 48 | 0.375 | 0.05 | 1.161 |
| MUS | 28 | 0.357 | 0.06 | 0.124 | 28 | 0.500 | 0.07 | 0.000 | 28 | 0.536 | 0.07 | 14.227 | 28 | 0.054 | 0.03 | 0.090 | 28 | 0.036 | 0.02 | 0.038 | 28 | 0.393 | 0.07 | 0.289 |
| MZO | 29 | 0.397 | 0.06 | 0.189 | 27 | 0.352 | 0.06 | 1.284 | 28 | 0.714 | 0.06 | 19.057 | 29 | 0.155 | 0.05 | 3.400 | 26 | 0.000 | 0.00 | 0.000 | 29 | 0.414 | 0.06 | 0.001 |
| PLN | 50 | 0.350 | 0.05 | 1.745 | 50 | 0.280 | 0.04 | 0.003 | 50 | 0.680 | 0.05 | 6.359 | 49 | 0.163 | 0.04 | 0.102 | 50 | 0.010 | 0.01 | 0.005 | 50 | 0.440 | 0.05 | 0.930 |
| RAJ | 52 | 0.462 | 0.05 | 0.361 | 52 | 0.298 | 0.04 | 0.169 | 52 | 0.692 | 0.05 | 7.044 | 52 | 0.115 | 0.03 | 0.885 | 52 | 0.067 | 0.02 | 2.851 | 51 | 0.392 | 0.05 | 3.439 |
| RIA | 50 | 0.380 | 0.05 | 2.785 | 50 | 0.350 | 0.05 | 18.265 | 49 | 0.653 | 0.05 | 40.570 | 50 | 0.200 | 0.04 | 0.000 | 51 | 0.000 | 0.00 | 0.000 | 50 | 0.330 | 0.05 | 0.081 |
| SAN | 24 | 0.479 | 0.07 | 1.526 | 21 | 0.262 | 0.07 | 3.099 | 24 | 0.708 | 0.07 | 3.744 | 23 | 0.239 | 0.06 | 0.616 | 20 | 0.025 | 0.02 | 0.013 | 24 | 0.500 | 0.07 | 0.667 |
| TAN | 16 | 0.531 | 0.09 | 2.315 | 15 | 0.233 | 0.08 | 1.389 | 16 | 0.594 | 0.09 | 5.980 | 16 | 0.437 | 0.09 | 1.165 | 16 | 0.031 | 0.03 | 0.017 | 15 | 0.333 | 0.09 | 0.600 |
| TOD | 49 | 0.531 | 0.05 | 2.572 | 49 | 0.194 | 0.04 | 0.592 | 50 | 0.700 | 0.05 | 13.719 | 50 | 0.000 | 0.00 | 0.000 | 50 | 0.050 | 0.02 | 6.787 | 48 | 0.594 | 0.05 | 1.323 |
| TRI | 50 | 0.400 | 0.05 | 0.000 | 51 | 0.216 | 0.04 | 0.095 | 49 | 0.745 | 0.04 | 13.083 | 50 | 0.170 | 0.04 | 0.199 | 51 | 0.010 | 0.01 | 0.005 | 51 | 0.461 | 0.05 | 0.218 |
| TTO | 30 | 0.433 | 0.06 | 0.222 | 29 | 0.086 | 0.04 | 0.258 | 30 | 0.667 | 0.06 | 14.700 | 30 | 0.150 | 0.05 | 3.600 | 30 | 0.000 | 0.00 | 0.000 | 30 | 0.967 | 0.02 | 0.036 |
| UBR | 27 | 0.407 | 0.07 | 1.395 | 27 | 0.407 | 0.07 | 7.702 | 27 | 0.630 | 0.07 | 7.400 | 27 | 0.111 | 0.04 | 1.688 | 27 | 0.130 | 0.05 | 0.599 | 27 | 0.370 | 0.07 | 1.977 |
| VAN | 50 | 0.380 | 0.05 | 0.219 | 50 | 0.310 | 0.05 | 0.624 | 50 | 0.590 | 0.05 | 14.866 | 50 | 0.140 | 0.03 | 1.435 | 50 | 0.030 | 0.02 | 0.048 | 50 | 0.430 | 0.05 | 1.678 |
| VLR | 43 | 0.337 | 0.05 | 2.074 | 40 | 0.412 | 0.06 | 2.049 | 40 | 0.712 | 0.05 | 13.126 | 43 | 0.186 | 0.04 | 2.317 | 43 | 0.023 | 0.02 | 0.024 | 43 | 0.395 | 0.05 | 0.212 |
| WBR | 23 | 0.391 | 0.07 | 0.175 | 23 | 0.391 | 0.07 | 1.675 | 23 | 0.478 | 0.07 | 9.763 | 23 | 0.130 | 0.05 | 0.518 | 23 | 0.152 | 0.05 | 0.741 | 22 | 0.455 | 0.08 | 0.220 |

| | Alu-PV92 | | | | MtNUC | | | | ESR1 | | | | NAT | | | | PSCR | | | | T2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ |
| AGH | 23 | 0.565 | 0.07 | 0.306 | 23 | 0.500 | 0.07 | 1.087 | 24 | 0.458 | 0.07 | 0.734 | 23 | 0.261 | 0.06 | 0.374 | 24 | 0.667 | 0.07 | 0.094 | 24 | 0.521 | 0.07 | 0.174 |
| AMB | 50 | 0.610 | 0.05 | 2.027 | 49 | 0.469 | 0.05 | 1.598 | 50 | 0.550 | 0.05 | 3.188 | 50 | 0.400 | 0.05 | 0.000 | 50 | 0.740 | 0.04 | 0.078 | 49 | 0.306 | 0.05 | 1.146 |
| BAG | 31 | 0.532 | 0.06 | 0.772 | 31 | 0.532 | 0.06 | 0.318 | 31 | 0.548 | 0.06 | 0.241 | 31 | 0.387 | 0.06 | 3.178 | 31 | 0.726 | 0.06 | 2.270 | 31 | 0.274 | 0.06 | 2.270 |
| CHA | 25 | 0.460 | 0.07 | 0.055 | 24 | 0.375 | 0.07 | 0.107 | 25 | 0.460 | 0.07 | 0.055 | 25 | 0.440 | 0.07 | 0.017 | 25 | 0.780 | 0.06 | 4.352 | 25 | 0.600 | 0.07 | 0.000 |
| CHK | 10 | 0.300 | 0.10 | 0.023 | 10 | 0.200 | 0.09 | 0.625 | 10 | 0.500 | 0.11 | 0.400 | 10 | 0.050 | 0.05 | 0.028 | 8 | 0.937 | 0.06 | 0.036 | 7 | 0.714 | 0.12 | 1.120 |
| GAU | 15 | 0.667 | 0.09 | 0.600 | 14 | 0.500 | 0.09 | 0.286 | 13 | 0.385 | 0.10 | 1.593 | 13 | 0.308 | 0.09 | 1.003 | 13 | 0.769 | 0.08 | 4.174 | 12 | 0.583 | 0.10 | 0.010 |
| HAL | 48 | 0.437 | 0.05 | 1.130 | 48 | 0.375 | 0.05 | 0.593 | 48 | 0.437 | 0.05 | 1.130 | 48 | 0.250 | 0.04 | 0.593 | 48 | 0.635 | 0.05 | 5.086 | 48 | 0.479 | 0.05 | 1.366 |
| HO | 54 | 0.435 | 0.05 | 0.016 | 53 | 0.547 | 0.05 | 0.232 | 54 | 0.315 | 0.04 | 0.167 | 54 | 0.259 | 0.04 | 0.069 | 53 | 0.736 | 0.04 | 0.046 | 53 | 0.368 | 0.05 | 1.163 |
| ILA | 49 | 0.551 | 0.05 | 0.257 | 47 | 0.426 | 0.05 | 0.790 | 49 | 0.327 | 0.05 | 0.021 | 50 | 0.240 | 0.04 | 4.986 | 50 | 0.670 | 0.05 | 0.126 | 50 | 0.320 | 0.05 | 1.898 |
| IYN | 51 | 0.539 | 0.05 | 0.009 | 51 | 0.382 | 0.05 | 4.199 | 51 | 0.412 | 0.05 | 0.140 | 51 | 0.294 | 0.05 | 0.907 | 51 | 0.539 | 0.05 | 1.498 | 50 | 0.370 | 0.05 | 1.253 |
| IYR | 50 | 0.570 | 0.05 | 1.025 | 48 | 0.417 | 0.05 | 0.039 | 50 | 0.470 | 0.05 | 0.352 | 49 | 0.286 | 0.05 | 0.000 | 50 | 0.540 | 0.05 | 2.158 | 50 | 0.430 | 0.05 | 1.678 |
| JAM | 55 | 0.209 | 0.04 | 0.236 | 53 | 0.708 | 0.04 | 5.297 | 55 | 0.382 | 0.05 | 0.338 | 55 | 0.100 | 0.03 | 4.719 | 55 | 0.982 | 0.01 | 0.019 | 53 | 0.623 | 0.05 | 2.217 |
| KMR | 56 | 0.446 | 0.05 | 4.310 | 54 | 0.537 | 0.05 | 0.609 | 57 | 0.500 | 0.05 | 0.439 | 57 | 0.123 | 0.03 | 1.966 | 57 | 0.772 | 0.04 | 0.527 | 55 | 0.345 | 0.05 | 2.337 |
| KOT | 45 | 0.700 | 0.05 | 4.385 | 45 | 0.433 | 0.05 | 0.075 | 45 | 0.667 | 0.05 | 0.450 | 45 | 0.189 | 0.04 | 0.147 | 45 | 0.756 | 0.05 | 1.858 | 45 | 0.233 | 0.04 | 0.141 |
| KUR | 54 | 0.287 | 0.04 | 0.134 | 53 | 0.462 | 0.05 | 0.032 | 54 | 0.241 | 0.04 | 0.707 | 54 | 0.111 | 0.03 | 0.844 | 54 | 0.648 | 0.05 | 1.908 | 54 | 0.241 | 0.04 | 0.707 |
| LOD | 32 | 0.484 | 0.06 | 1.139 | 32 | 0.562 | 0.06 | 1.814 | 32 | 0.203 | 0.05 | 2.079 | 31 | 0.323 | 0.06 | 1.015 | 32 | 0.594 | 0.06 | 0.277 | 32 | 0.297 | 0.06 | 0.023 |
| MAH | 34 | 0.485 | 0.06 | 0.478 | 34 | 0.500 | 0.06 | 0.118 | 34 | 0.426 | 0.06 | 0.017 | 34 | 0.368 | 0.06 | 1.073 | 34 | 0.779 | 0.05 | 0.119 | 34 | 0.412 | 0.06 | 0.028 |
| MOG | 25 | 0.300 | 0.06 | 0.057 | 22 | 0.432 | 0.07 | 0.008 | 23 | 0.543 | 0.07 | 3.439 | 25 | 0.220 | 0.06 | 1.989 | 25 | 0.940 | 0.03 | 0.102 | 23 | 0.587 | 0.07 | 0.857 |
| MUR | 49 | 0.480 | 0.05 | 0.980 | 49 | 0.612 | 0.05 | 0.145 | 49 | 0.480 | 0.05 | 0.980 | 49 | 0.153 | 0.04 | 0.882 | 49 | 0.714 | 0.05 | 0.490 | 49 | 0.520 | 0.05 | 0.980 |
| MUS | 28 | 0.679 | 0.06 | 3.333 | 27 | 0.537 | 0.07 | 0.882 | 28 | 0.393 | 0.07 | 0.065 | 28 | 0.357 | 0.06 | 1.383 | 28 | 0.446 | 0.07 | 0.197 | 28 | 0.500 | 0.07 | 0.000 |
| MZO | 29 | 0.138 | 0.05 | 0.742 | 27 | 0.574 | 0.07 | 0.006 | 29 | 0.276 | 0.06 | 4.209 | 26 | 0.000 | 0.00 | 0.000 | 26 | 1.000 | 0.00 | 0.000 | 26 | 0.538 | 0.07 | 4.013 |
| PLN | 50 | 0.560 | 0.05 | 6.148 | 47 | 0.564 | 0.05 | 1.326 | 50 | 0.470 | 0.05 | 0.294 | 50 | 0.270 | 0.04 | 0.945 | 50 | 0.720 | 0.04 | 0.574 | 50 | 0.430 | 0.05 | 0.516 |
| RAJ | 52 | 0.663 | 0.05 | 1.718 | 51 | 0.529 | 0.05 | 0.157 | 51 | 0.373 | 0.05 | 3.401 | 50 | 0.300 | 0.05 | 0.113 | 50 | 0.710 | 0.05 | 4.846 | 49 | 0.500 | 0.05 | 0.020 |
| RIA | 50 | 0.260 | 0.04 | 0.208 | 50 | 0.660 | 0.05 | 0.591 | 50 | 0.410 | 0.05 | 0.870 | 50 | 0.080 | 0.03 | 0.378 | 49 | 0.990 | 0.01 | 0.005 | 44 | 0.580 | 0.05 | 0.232 |
| SAN | 24 | 0.437 | 0.07 | 3.984 | 23 | 0.522 | 0.07 | 0.048 | 20 | 0.250 | 0.07 | 2.222 | 23 | 0.304 | 0.07 | 4.402 | 22 | 0.659 | 0.07 | 0.279 | 24 | 0.458 | 0.07 | 5.916 |
| TAN | 16 | 0.344 | 0.08 | 0.015 | 14 | 0.464 | 0.09 | 1.196 | 16 | 0.437 | 0.09 | 0.004 | 16 | 0.344 | 0.08 | 0.974 | 15 | 0.433 | 0.09 | 0.037 | 14 | 0.500 | 0.09 | 0.286 |
| TOD | 49 | 0.745 | 0.04 | 0.020 | 49 | 0.796 | 0.04 | 2.969 | 48 | 0.604 | 0.05 | 0.099 | 48 | 0.312 | 0.05 | 0.044 | 46 | 0.793 | 0.04 | 3.116 | 49 | 0.112 | 0.03 | 0.783 |
| TRI | 49 | 0.194 | 0.04 | 1.121 | 50 | 0.600 | 0.05 | 1.389 | 51 | 0.402 | 0.05 | 0.196 | 51 | 0.108 | 0.03 | 0.351 | 50 | 0.950 | 0.02 | 0.139 | 51 | 0.716 | 0.04 | 0.365 |
| TTO | 30 | 0.067 | 0.03 | 0.153 | 30 | 0.317 | 0.06 | 0.700 | 30 | 0.617 | 0.06 | 0.099 | 30 | 0.033 | 0.02 | 0.036 | 30 | 0.950 | 0.03 | 0.083 | 30 | 0.600 | 0.06 | 1.875 |
| UBR | 27 | 0.667 | 0.06 | 0.750 | 27 | 0.556 | 0.07 | 0.270 | 27 | 0.222 | 0.06 | 2.204 | 27 | 0.241 | 0.06 | 2.714 | 27 | 0.500 | 0.07 | 0.926 | 27 | 0.481 | 0.07 | 0.942 |
| VAN | 50 | 0.510 | 0.05 | 0.000 | 50 | 0.510 | 0.05 | 1.287 | 50 | 0.440 | 0.05 | 0.574 | 50 | 0.290 | 0.05 | 0.020 | 50 | 0.650 | 0.05 | 0.006 | 48 | 0.437 | 0.05 | 0.485 |
| VLR | 42 | 0.512 | 0.05 | 0.377 | 43 | 0.314 | 0.05 | 0.769 | 41 | 0.402 | 0.05 | 0.055 | 40 | 0.225 | 0.05 | 0.001 | 39 | 0.603 | 0.06 | 1.513 | 42 | 0.369 | 0.05 | 0.034 |
| WBR | 23 | 0.435 | 0.07 | 1.308 | 23 | 0.304 | 0.07 | 1.239 | 23 | 0.370 | 0.07 | 2.767 | 23 | 0.413 | 0.07 | 6.324 | 23 | 0.609 | 0.07 | 1.675 | 23 | 0.587 | 0.07 | 2.738 |

| | LPL | | | | ALB | | | | CYP1A-MspI | | | | HOXB4-MspI | | | | ADH-RsaI | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ | *n* | *p* | *S.E.* | $\chi^2$ |
| AGH | 24 | 0.375 | 0.07 | 0.107 | 23 | 0.522 | 0.07 | 0.381 | 24 | 0.729 | 0.06 | 0.618 | 24 | 0.521 | 0.07 | 0.174 | 22 | 0.568 | 0.07 | 0.008 |
| AMB | 50 | 0.350 | 0.05 | 0.006 | 50 | 0.540 | 0.05 | 0.109 | 50 | 0.500 | 0.05 | 2.000 | 50 | 0.500 | 0.05 | 0.720 | 50 | 0.640 | 0.05 | 0.087 |
| BAG | 31 | 0.339 | 0.06 | 4.202 | 31 | 0.532 | 0.06 | 1.653 | 31 | 0.548 | 0.06 | 0.920 | 31 | 0.452 | 0.06 | 2.837 | 31 | 0.355 | 0.06 | 0.740 |
| CHA | 25 | 0.360 | 0.07 | 0.043 | 25 | 0.380 | 0.07 | 4.909 | 25 | 0.440 | 0.07 | 3.074 | 25 | 0.580 | 0.07 | 1.340 | 24 | 0.562 | 0.07 | 0.243 |
| CHK | 10 | 0.600 | 0.11 | 0.278 | 10 | 0.650 | 0.11 | 1.160 | 10 | 0.450 | 0.11 | 1.715 | 10 | 0.800 | 0.09 | 0.625 | 9 | 0.778 | 0.10 | 1.148 |
| GAU | 13 | 0.423 | 0.10 | 0.138 | 13 | 0.346 | 0.09 | 0.294 | 12 | 0.458 | 0.10 | 0.367 | 13 | 0.462 | 0.10 | 1.887 | 13 | 0.538 | 0.10 | 0.066 |
| HAL | 48 | 0.417 | 0.05 | 0.157 | 48 | 0.417 | 0.05 | 0.039 | 46 | 0.565 | 0.05 | 0.612 | 47 | 0.383 | 0.05 | 0.304 | 47 | 0.638 | 0.05 | 0.289 |
| HO | 54 | 0.426 | 0.05 | 0.013 | 54 | 0.667 | 0.05 | 0.000 | 54 | 0.528 | 0.05 | 0.274 | 54 | 0.361 | 0.05 | 0.378 | 54 | 0.602 | 0.05 | 0.062 |
| ILA | 50 | 0.460 | 0.05 | 0.809 | 49 | 0.602 | 0.05 | 0.205 | 49 | 0.388 | 0.05 | 0.677 | 49 | 0.490 | 0.05 | 0.020 | 46 | 0.641 | 0.05 | 0.347 |
| IYN | 51 | 0.402 | 0.05 | 0.522 | 51 | 0.402 | 0.05 | 1.051 | 51 | 0.716 | 0.04 | 0.365 | 51 | 0.588 | 0.05 | 0.042 | 51 | 0.569 | 0.05 | 0.743 |
| IYR | 50 | 0.370 | 0.05 | 0.263 | 50 | 0.470 | 0.05 | 2.814 | 50 | 0.660 | 0.05 | 0.019 | 50 | 0.500 | 0.05 | 0.080 | 50 | 0.540 | 0.05 | 1.898 |
| JAM | 54 | 0.333 | 0.05 | 0.000 | 52 | 0.433 | 0.05 | 0.022 | 55 | 0.518 | 0.05 | 3.046 | 54 | 0.648 | 0.05 | 0.616 | 55 | 0.400 | 0.05 | 0.455 |
| KMR | 57 | 0.360 | 0.04 | 0.046 | 54 | 0.546 | 0.05 | 1.070 | 56 | 0.402 | 0.05 | 0.000 | 57 | 0.289 | 0.04 | 4.310 | 51 | 0.529 | 0.05 | 3.428 |
| KOT | 45 | 0.211 | 0.04 | 3.223 | 45 | 0.256 | 0.05 | 2.608 | 45 | 0.578 | 0.05 | 1.461 | 45 | 0.444 | 0.05 | 0.450 | 45 | 0.711 | 0.05 | 3.998 |
| KUR | 54 | 0.269 | 0.04 | 0.005 | 53 | 0.623 | 0.05 | 0.102 | 54 | 0.583 | 0.05 | 1.768 | 52 | 0.596 | 0.05 | 2.106 | 51 | 0.686 | 0.05 | 1.659 |
| LOD | 32 | 0.547 | 0.06 | 5.986 | 32 | 0.437 | 0.06 | 5.039 | 32 | 0.594 | 0.06 | 0.042 | 32 | 0.437 | 0.06 | 0.008 | 32 | 0.594 | 0.06 | 0.277 |
| MAH | 34 | 0.279 | 0.05 | 0.311 | 34 | 0.485 | 0.06 | 1.900 | 34 | 0.721 | 0.05 | 1.314 | 34 | 0.485 | 0.06 | 0.478 | 34 | 0.441 | 0.06 | 0.185 |
| MOG | 25 | 0.380 | 0.07 | 0.268 | 25 | 0.560 | 0.07 | 0.887 | 25 | 0.460 | 0.07 | 0.055 | 23 | 0.609 | 0.07 | 0.209 | 24 | 0.479 | 0.07 | 8.146 |
| MUR | 49 | 0.449 | 0.05 | 0.257 | 49 | 0.490 | 0.05 | 1.007 | 47 | 0.468 | 0.05 | 2.506 | 48 | 0.323 | 0.05 | 0.440 | 49 | 0.735 | 0.04 | 0.163 |
| MUS | 28 | 0.375 | 0.06 | 0.571 | 27 | 0.426 | 0.07 | 0.006 | 28 | 0.821 | 0.05 | 2.035 | 28 | 0.500 | 0.07 | 0.000 | 28 | 0.500 | 0.07 | 0.571 |
| MZO | 26 | 0.385 | 0.07 | 0.016 | 29 | 0.517 | 0.07 | 2.778 | 26 | 0.692 | 0.06 | 0.246 | 25 | 0.620 | 0.07 | 1.392 | 26 | 0.615 | 0.07 | 6.829 |
| PLN | 48 | 0.437 | 0.05 | 0.227 | 49 | 0.541 | 0.05 | 0.036 | 49 | 0.765 | 0.04 | 0.309 | 50 | 0.520 | 0.05 | 0.742 | 49 | 0.520 | 0.05 | 0.024 |
| RAJ | 51 | 0.461 | 0.05 | 0.009 | 49 | 0.398 | 0.05 | 0.020 | 51 | 0.647 | 0.05 | 2.081 | 51 | 0.657 | 0.05 | 0.382 | 48 | 0.542 | 0.05 | 1.467 |
| RIA | 48 | 0.365 | 0.05 | 0.739 | 41 | 0.439 | 0.05 | 3.858 | 41 | 0.524 | 0.06 | 0.206 | 36 | 0.694 | 0.05 | 0.080 | 46 | 0.467 | 0.05 | 0.317 |
| SAN | 25 | 0.480 | 0.07 | 0.371 | 20 | 0.450 | 0.08 | 3.430 | 20 | 0.450 | 0.08 | 3.430 | 20 | 0.175 | 0.06 | 0.900 | 20 | 0.725 | 0.07 | 0.299 |
| TAN | 16 | 0.375 | 0.09 | 0.071 | 15 | 0.533 | 0.09 | 0.077 | 15 | 0.533 | 0.09 | 0.579 | 16 | 0.469 | 0.09 | 0.236 | 15 | 0.600 | 0.09 | 0.185 |
| TOD | 47 | 0.564 | 0.05 | 0.312 | 45 | 0.467 | 0.05 | 1.736 | 49 | 0.816 | 0.04 | 2.481 | 36 | 0.472 | 0.06 | 0.472 | 47 | 0.447 | 0.05 | 0.133 |
| TRI | 50 | 0.420 | 0.05 | 4.918 | 49 | 0.510 | 0.05 | 0.506 | 50 | 0.560 | 0.05 | 0.034 | 50 | 0.690 | 0.05 | 1.424 | 50 | 0.480 | 0.05 | 0.703 |
| TTO | 30 | 0.300 | 0.06 | 2.184 | 30 | 0.367 | 0.06 | 0.660 | 30 | 0.567 | 0.06 | 1.033 | 30 | 0.567 | 0.06 | 0.074 | 30 | 0.417 | 0.06 | 1.811 |
| UBR | 27 | 0.278 | 0.06 | 0.773 | 27 | 0.389 | 0.07 | 2.846 | 27 | 0.611 | 0.07 | 0.770 | 27 | 0.648 | 0.06 | 0.308 | 27 | 0.611 | 0.07 | 0.551 |
| VAN | 50 | 0.360 | 0.05 | 0.870 | 50 | 0.450 | 0.05 | 7.759 | 50 | 0.580 | 0.05 | 0.227 | 50 | 0.490 | 0.05 | 0.323 | 50 | 0.430 | 0.05 | 0.516 |
| VLR | 42 | 0.357 | 0.05 | 5.091 | 37 | 0.541 | 0.06 | 0.620 | 41 | 0.610 | 0.05 | 0.246 | 36 | 0.514 | 0.06 | 0.992 | 42 | 0.357 | 0.05 | 0.832 |
| WBR | 23 | 0.413 | 0.07 | 0.004 | 23 | 0.391 | 0.07 | 1.775 | 23 | 0.717 | 0.07 | 4.949 | 22 | 0.545 | 0.08 | 0.220 | 22 | 0.341 | 0.07 | 0.279 |

Table 2.14
Sample Sizes and Haplotype Frequencies at 3 Linked Loci in Ethnic Populations of India

| Population | n | ALAD: MspI - RsaI | | | | DRD2: Taq "A" - Taq "D" - Taq "B" | | | | | | | | HB: $\psi\beta$ *Hinc II* - 3'$\psi\beta$ *HincII* - 5' $\beta$ *HinfI* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ++ | +- | -+ | -- | +++ | ++- | +-+ | +-- | -++ | -+- | --+ | --- | +++ | ++- | +-+ | +-- | -++ | -+- | --+ | --- |
| AGH | 24 | 0.001 | 0.195 | 0.151 | 0.653 | 0.171 | 0.079 | 0.205 | 0 | 0.125 | 0.421 | 0 | 0 | 0.047 | 0.307 | 0 | 0 | 0.207 | 0.022 | 0.35 | 0.066 |
| AMB | 50 | 0 | 0.16 | 0.11 | 0.73 | 0.537 | 0.043 | 0.26 | 0.02 | 0.013 | 0.087 | 0.039 | 0.001 | 0.163 | 0.007 | 0 | 0 | 0.24 | 0 | 0.547 | 0.043 |
| BAG | 31 | 0.001 | 0.144 | 0.16 | 0.694 | 0.305 | 0.034 | 0.258 | 0 | 0.082 | 0.321 | 0 | 0 | 0.279 | 0 | 0.038 | 0.021 | 0.124 | 0 | 0.397 | 0.14 |
| CHA | 25 | 0.072 | 0.028 | 0.078 | 0.822 | 0.283 | 0.077 | 0.32 | 0 | 0.097 | 0.223 | 0 | 0 | 0.19 | 0.03 | 0 | 0 | 0.1 | 0.02 | 0.57 | 0.09 |
| CHK | 10 | 0 | 0 | 0.35 | 0.65 | 0.65 | 0 | 0 | 0 | 0 | 0.35 | 0 | 0 | 0.45 | 0 | 0.15 | 0 | 0 | 0 | 0.4 | 0 |
| GAU | 15 | 0 | 0.208 | 0.208 | 0.583 | 0.417 | 0 | 0.25 | 0 | 0 | 0.333 | 0 | 0 | 0.154 | 0.279 | 0 | 0 | 0.033 | 0.033 | 0.412 | 0.088 |
| HAL | 48 | 0 | 0.128 | 0.256 | 0.616 | 0.337 | 0.012 | 0.128 | 0.012 | 0.125 | 0.371 | 0 | 0.015 | 0.192 | 0.142 | 0 | 0 | 0.178 | 0.061 | 0.39 | 0.037 |
| HO | 49 | 0 | 0.111 | 0.315 | 0.574 | 0.211 | 0.039 | 0.163 | 0 | 0.126 | 0.461 | 0 | 0 | 0.13 | 0.105 | 0.026 | 0.025 | 0.132 | 0 | 0.569 | 0.013 |
| ILA | 50 | 0.006 | 0.218 | 0.157 | 0.619 | 0.162 | 0.042 | 0.429 | 0 | 0.021 | 0.346 | 0 | 0 | 0.162 | 0.058 | 0 | 0 | 0.29 | 0.03 | 0.447 | 0.012 |
| IYN | 51 | 0 | 0.167 | 0.275 | 0.558 | 0.356 | 0.01 | 0.388 | 0 | 0.056 | 0.176 | 0.014 | 0 | 0.263 | 0.039 | 0.012 | 0 | 0.12 | 0.03 | 0.537 | 0 |
| IYR | 50 | 0 | 0.15 | 0.24 | 0.61 | 0.337 | 0.023 | 0.432 | 0.017 | 0.02 | 0.159 | 0.01 | 0 | 0.234 | 0.046 | 0 | 0 | 0.151 | 0.029 | 0.525 | 0.015 |
| JAM | 55 | 0 | 0.065 | 0.407 | 0.528 | 0.718 | 0.009 | 0.036 | 0 | 0.028 | 0.209 | 0 | 0 | 0.171 | 0 | 0.029 | 0 | 0.347 | 0 | 0.38 | 0.073 |
| KMR | 57 | 0 | 0.228 | 0.368 | 0.404 | 0.234 | 0.038 | 0.246 | 0.019 | 0.2 | 0.235 | 0.03 | 0 | 0.209 | 0.103 | 0.013 | 0 | 0.221 | 0.09 | 0.337 | 0.027 |
| KOT | 45 | 0 | 0.233 | 0.178 | 0.589 | 0.441 | 0 | 0.048 | 0 | 0.059 | 0.233 | 0.219 | 0 | 0.311 | 0 | 0 | 0 | 0.122 | 0.011 | 0.556 | 0 |
| KUR | 54 | 0 | 0.222 | 0.482 | 0.296 | 0.145 | 0 | 0.381 | 0.011 | 0.08 | 0.275 | 0.06 | 0.048 | 0.154 | 0.272 | 0 | 0 | 0.133 | 0.006 | 0.417 | 0.018 |
| LOD | 32 | 0 | 0.25 | 0.422 | 0.328 | 0.292 | 0 | 0.159 | 0.018 | 0.194 | 0.279 | 0.058 | 0 | 0.081 | 0.091 | 0 | 0 | 0.049 | 0.013 | 0.729 | 0.036 |
| MAH | 34 | 0.004 | 0.132 | 0.208 | 0.655 | 0.412 | 0 | 0.235 | 0.029 | 0 | 0.309 | 0.015 | 0 | 0.082 | 0.059 | 0.065 | 0 | 0.079 | 0 | 0.715 | 0 |
| MOG | 25 | 0 | 0.023 | 0.341 | 0.636 | 0.575 | 0 | 0.075 | 0 | 0 | 0.35 | 0 | 0 | 0.439 | 0.02 | 0.021 | 0 | 0.046 | 0.015 | 0.414 | 0.045 |
| MUR | 49 | 0 | 0.202 | 0.308 | 0.489 | 0.354 | 0 | 0.146 | 0 | 0.073 | 0.427 | 0 | 0 | 0.163 | 0.133 | 0 | 0 | 0.209 | 0.056 | 0.424 | 0.015 |
| MUS | 28 | 0 | 0.161 | 0.125 | 0.714 | 0.446 | 0 | 0.375 | 0 | 0 | 0.143 | 0 | 0 | 0.25 | 0.054 | 0 | 0 | 0.161 | 0 | 0.536 | 0 |
| MZO | 25 | 0 | 0.071 | 0.214 | 0.715 | 0.731 | 0 | 0 | 0 | 0.019 | 0.231 | 0 | 0.019 | 0.04 | 0 | 0 | 0 | 0.18 | 0 | 0.76 | 0.02 |
| PLN | 49 | 0 | 0.14 | 0.28 | 0.58 | 0.37 | 0.66 | 0.213 | 0 | 0.045 | 0.306 | 0 | 0 | 0.265 | 0.039 | 0.023 | 0 | 0.14 | 0.006 | 0.512 | 0.017 |
| RAJ | 52 | 0 | 0.098 | 0.333 | 0.569 | 0.407 | 0.022 | 0.312 | 0 | 0.023 | 0.198 | 0.031 | 0 | 0.266 | 0.079 | 0.011 | 0 | 0.078 | 0 | 0.549 | 0.017 |
| RIA | 48 | 0.087 | 0.076 | 0.313 | 0.524 | 0.503 | 0 | 0.148 | 0.012 | 0.023 | 0.253 | 0 | 0.061 | 0.593 | 0.035 | 0.094 | 0.038 | 0.069 | 0 | 0.17 | 0 |
| SAN | 25 | 0 | 0.2 | 0.2 | 0.6 | 0.35 | 0 | 0.225 | 0 | 0.075 | 0.35 | 0 | 0 | 0.08 | 0.04 | 0 | 0 | 0.26 | 0.08 | 0.54 | 0 |
| TAN | 16 | 0.067 | 0.133 | 0.267 | 0.533 | 0.26 | 0.106 | 0.1 | 0 | 0.073 | 0.461 | 0 | 0 | 0.068 | 0.368 | 0.033 | 0 | 0.12 | 0.038 | 0.373 | 0 |
| TOD | 48 | 0 | 0.109 | 0.261 | 0.63 | 0.292 | 0 | 0.487 | 0 | 0.046 | 0.093 | 0.082 | 0 | 0.147 | 0.014 | 0.088 | 0 | 0.12 | 0 | 0.603 | 0.028 |
| TRI | 51 | 0 | 0.076 | 0.359 | 0.565 | 0.459 | 0 | 0.143 | 0 | 0.035 | 0.272 | 0.016 | 0.075 | 0.338 | 0 | 0.034 | 0 | 0.083 | 0 | 0.544 | 0 |
| TTO | 30 | 0 | 0.033 | 0.633 | 0.333 | 0.201 | 0 | 0.165 | 0 | 0.017 | 0.432 | 0 | 0.185 | 0.3 | 0 | 0 | 0 | 0.05 | 0 | 0.65 | 0 |
| UBR | 27 | 0 | 0.13 | 0.37 | 0.5 | 0.277 | 0.02 | 0.37 | 0 | 0.057 | 0.276 | 0 | 0 | 0.199 | 0 | 0.023 | 0.019 | 0.174 | 0.053 | 0.474 | 0.059 |
| VAN | 50 | 0 | 0.15 | 0.18 | 0.67 | 0.399 | 0.064 | 0.158 | 0 | 0.041 | 0.316 | 0.022 | 0 | 0.121 | 0.049 | 0 | 0 | 0.241 | 0.039 | 0.508 | 0.042 |
| VLR | 42 | 0.001 | 0.187 | 0.287 | 0.525 | 0.318 | 0.026 | 0.331 | 0.012 | 0.02 | 0.274 | 0.02 | 0 | 0.248 | 0.041 | 0.043 | 0 | 0.18 | 0.029 | 0.457 | 0 |
| WBR | 23 | 0 | 0.159 | 0.159 | 0.682 | 0.36 | 0 | 0.345 | 0 | 0.095 | 0.136 | 0.064 | 0 | 0.142 | 0.054 | 0 | 0 | 0.068 | 0.062 | 0.595 | 0.079 |

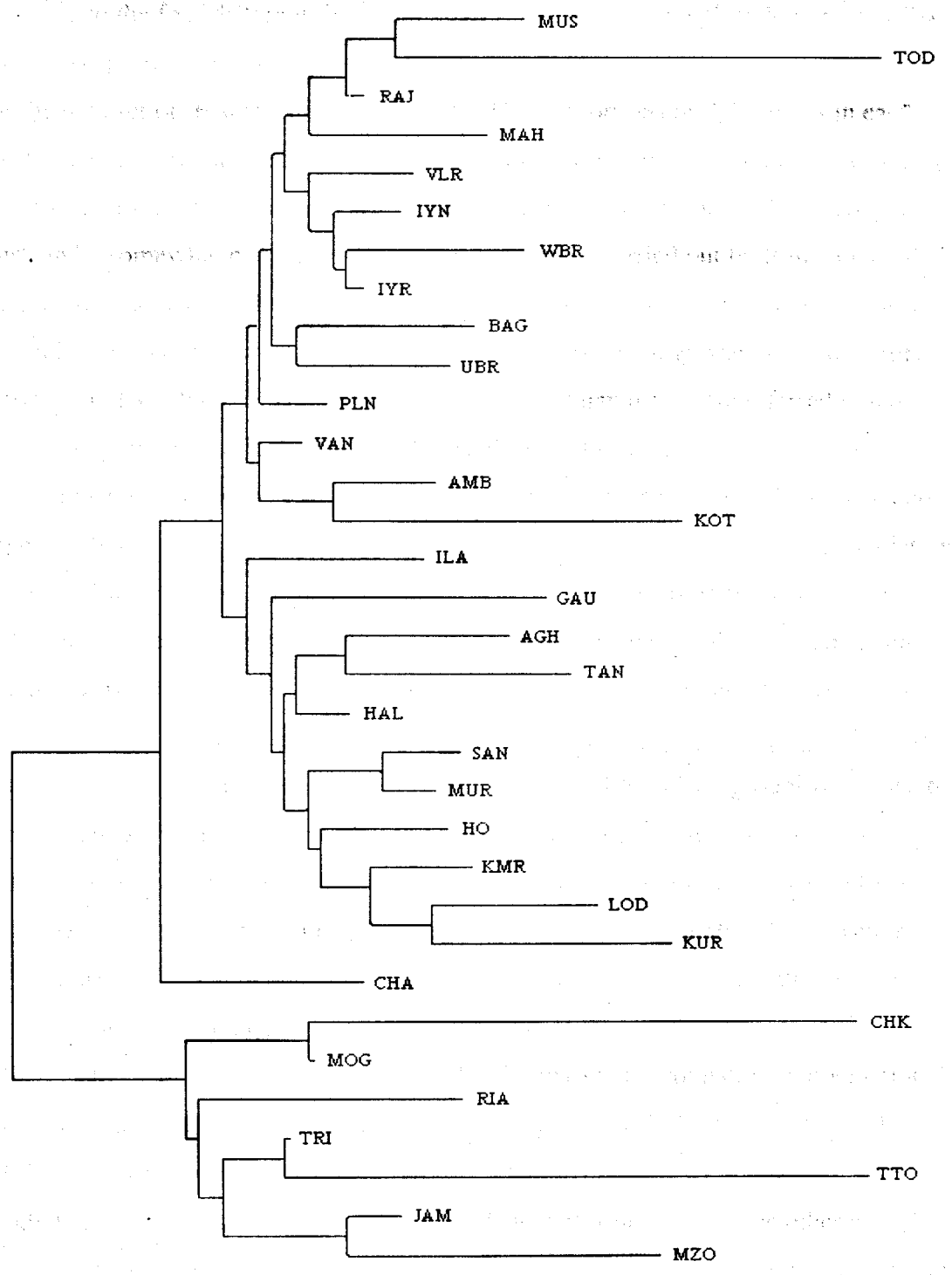



**Figure 2.16**
Neighbor-joining tree depicting genetic affinities among Indian ethnic populations based on frequencies of autosomal unlinked IDP and RSP markers and haplotype frequencies of three sets of linked RSP markers

tribals belonging to the four different linguistic groups, and (b) individuals belonging to tribal, caste and religious groups cross-classified by linguistic affinity.

For the data set (a) K was estimated to be 5. The proportions of individuals in each subgroup estimated to originate in the "hypothetical" populations are depicted in Figures 2.17 and 2.18 for the data sets (a) and (b), respectively. The estimate of the number of "hypothetical" populations can be somewhat crude, but extensive simulations carried out by Pritchard et al. (2000) indicate that the results of this analysis can provide a fairly good idea of genetic structure. Figure 2.17 indicates that the Tibeto-Burman tribals are clearly distinguished from the other groups of tribals in that a large fraction (41%) of Tibeto-Burman tribals are inferred to have originated from population 4, while the proportions of individuals originating from this population are much smaller (11-14%). Conversely, while the Austro-Asiatic, Dravidian and Indo-European tribals have high proportions (23-31%) of individuals originating in populations 1 and 5, the Tibeto-Burman tribals have a low proportion (about 10%) of individuals originating in these populations. In general, this analysis indicates that with respect to the autosomal genes, the Austro-Asiatic, Dravidian and Indo-European tribals are more similar to one another than the Tibeto-Burman tribals.

Data set (b) comprised 11 subgroups of populations -- 4 linguistic groups of tribals, 6 groups of castes cross-classified by social rank and language, and 1 religious group (Muslim). For this data set, K was estimated to be 7. The estimated proportions of individuals belonging to the various subgroups originating from the various "hypothetical" populations are given in Figure 2.18. Some remarkable features are discernible from this figure. First, the caste and tribal populations are, by and large, distinct. The primary distinctive features are that while the caste populations have low proportions (about 12%) of individuals originating in population 3, these proportions are higher (about 20%) for the tribal populations. The contributions of population 5 to the caste populations is generally much higher than tribal populations. Second, as has already been noted, with the exception of the Tibeto-Burman tribals, the other groups of tribals are largely similar. Third, the Dravidian upper castes show a greater similarity with Indo-European upper castes, while the Dravidian lower castes are largely similar to the Dravidian tribals. Fourth, the Indo-European tribal group (Halba) is very similar to the Dravdian tribals, indicating that the Halbas may have been a Dravidian-speaking population and speak an Indo-European language possibly because of linguistic dominance. Fifth, the Muslims show greater

**Figure 2.17**
Results of population structure analysis for linguistic subgroups of tribals: Proportions of individuals originating in each inferred population

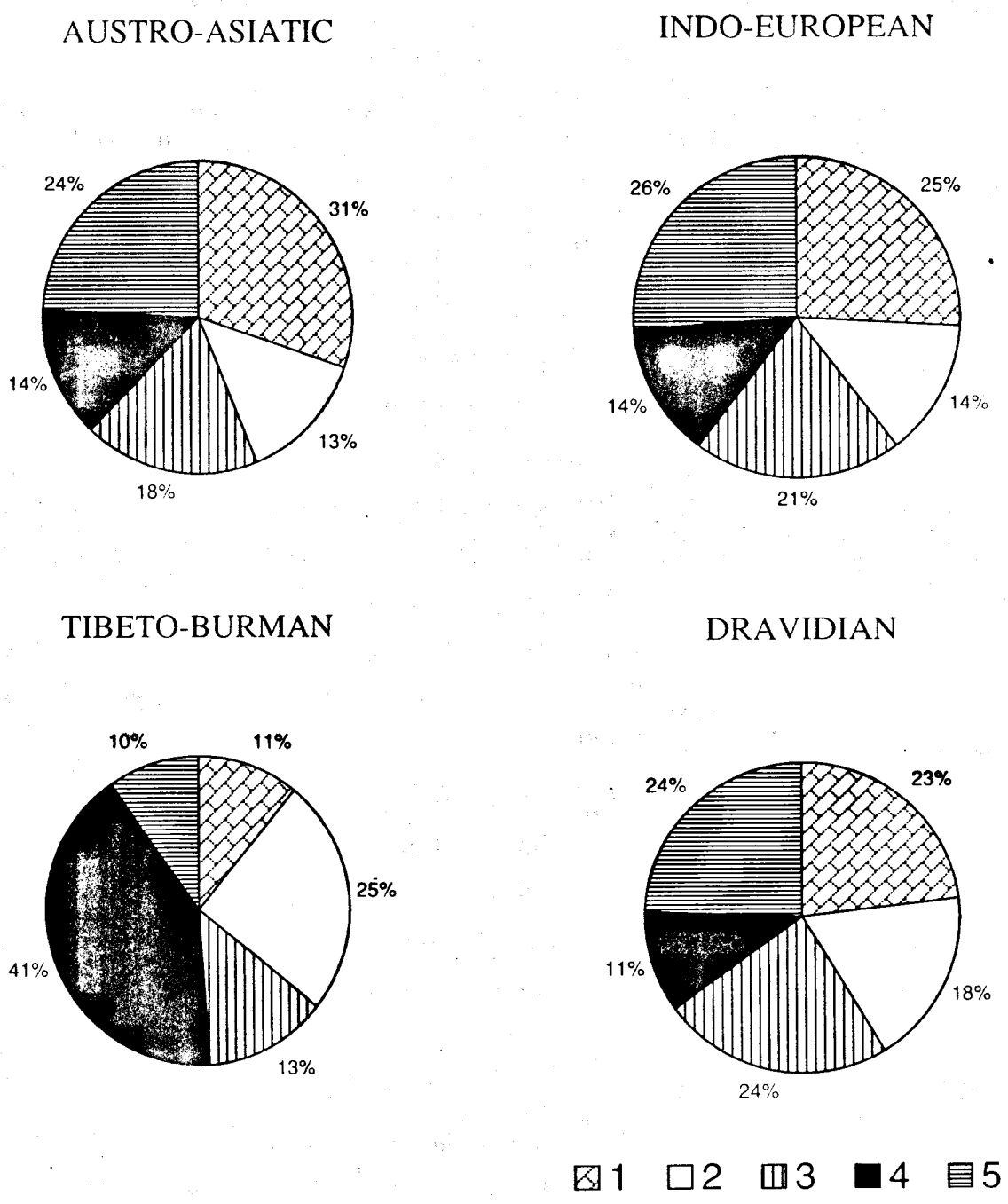
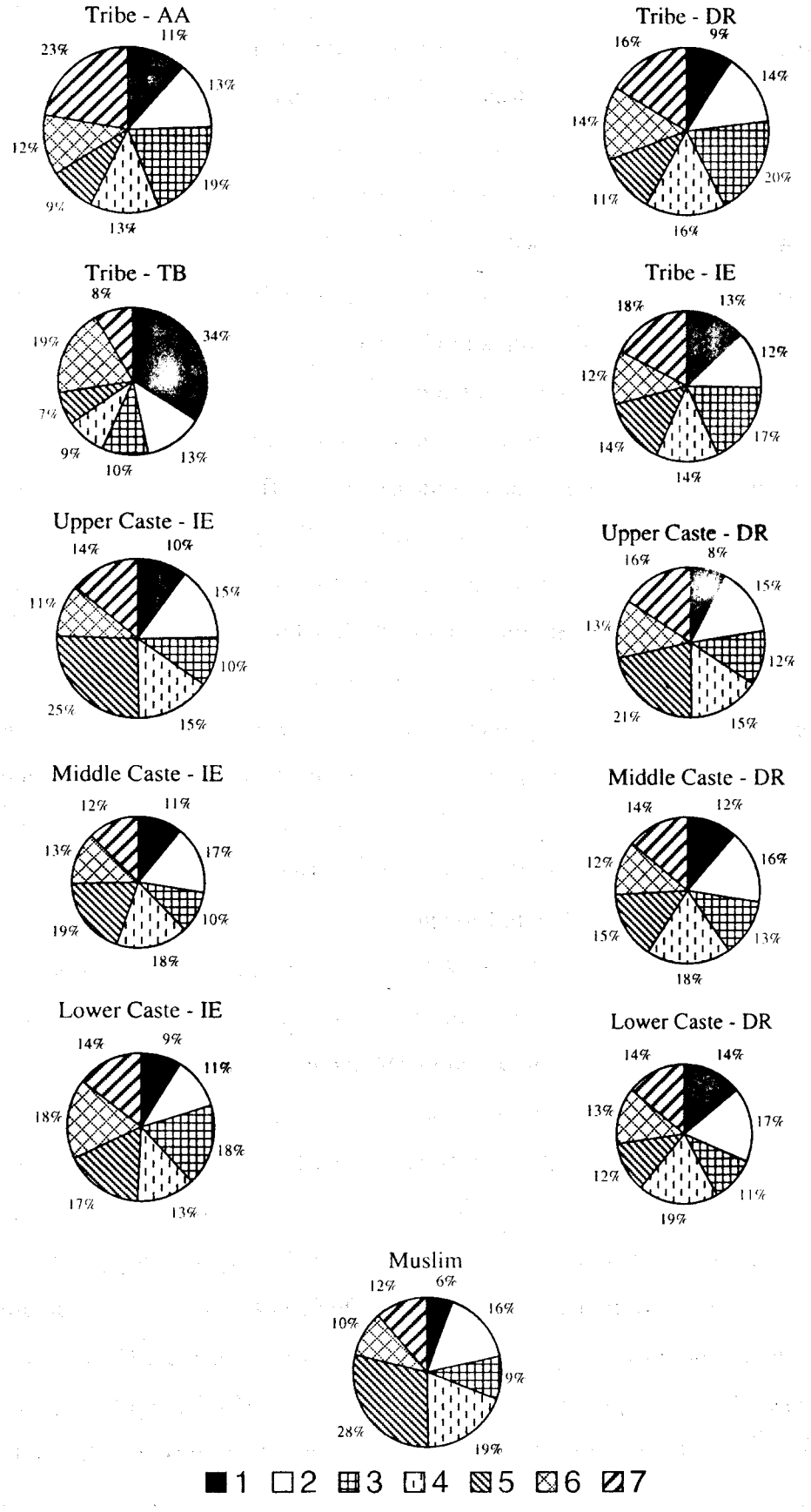


**Figure 2.18** Results of population structure analysis for various linguistic x socio-cultural subgroups: Proportions of individuals originating in each inferred population



similarities with Indo-European castes than with Dravidian castes or tribals. This is not surprising because it is known that in northern regions of India there have been conversions to Islam in historical times, predominantly from caste groups.

## Discussion

The present study provides the first comprehensive genomic view of the ethnic populations of India. Although many earlier studies using different sets of DNA markers have included samples from some ethnic populations of India, none of these studies has provided a comprehensive genomic view of ethnic populations of India. The primary reason has been the lack of systematic sampling of populations to ensure a wide geographical and socio-cultural coverage of ethnic India. Because past studies based on genetic markers (reviewed in Cavalli-Sforza et al. 1994; Majumder 1998) have shown that there is enormous genetic diversity among ethnic populations inhabiting different geographical regions and belonging to different socio-cultural subgroups, we considered it crucial to sample ethnic populations of India with a wide geographical and socio-cultural coverage and to generate data on a uniform set of DNA markers in order to obtain a holistic view the genomic composition of ethnic India and to draw inferences on the process of peopling of this region, which is considered to have been central in human evolution and dispersal (Cann 2001).

It has been argued (Risley 1915; Thapar 1966; Pattanayak 1998) that the Austro-Asiatic speaking tribals are the original inhabitants of India. Some other scholars have, however, argued that tribal groups speaking Dravidian and Austro-Asiatic languages have evolved from an older original substrate of proto-Australoids (Keith 1936), while the Tibeto-Burman speaking tribals are later immigrants from Tibet and Myanmar (Guha 1935). Parpola (1975) has contended that the different language families in India may represent different lineages, which is consistent with Cavalli-Sforza et al.'s (1994) finding that within India, linguistic differences account for much of the genetic diversity.

We have found that there is an underlying unity of female lineages in India as evidenced by the nearly uniform existence of a modal mtDNA haplotype across ethnic populations of India (Table 2.2). There is also no strong clustering of populations based on mtDNA RSP haplotype frequencies (Figure 2.10) or of the HVS1 sequences by population. However, there are two notable differences in the frequencies of this modal haplotype. First, the tribal populations show

significantly higher frequencies of this modal haplotype compared to the castes. As can be reasonably assumed from its distribution across geographical regions and socio-cultural groups, this modal haplotype is the most ancient in India. Therefore, its higher frequency among tribal populations is a testimony to these their being more ancient than the caste populations. Further, the coalescent times of the major subclades of M (e.g., M2) found in India are greater than most east Asian and Papuan branches of haplogroup M (Forster et al. 2001), suggesting that India was settled early after humankind came out of Africa (Kivisild et al. 1999b). The extent of genetic differentiation is also higher in respect of all three sets of markers (mt, Y and autosomal) among tribal groups than among caste groups (Table 2.7), indicating that the tribal groups are not only ancient but have remained isolated. The Y-haplogroup frequency profiles of castes and tribes are significantly different. There is also virtually no sharing of Y-STRP haplotypes between tribes and castes. Second, among the tribal groups, the Austro-Asiatic tribals of eastern India possess the highest frequencies of the modal haplotype, which is consistent with the notion (Risley 1915; Thapar 1966; Pattanayak 1998) that the Austro-Asiatic tribals may the most ancient in India. In fact, this group also possesses the highest frequency of haplogroup M (the background on which the modal haplotype occurs), which is supposedly an an ancient east Asian marker (Ballinger et al. 1992). The clade M2, which is the most frequent clade of M and has the highest nucleotide diversity compared to the other clades of M found in India and therefore is likely to be the most ancient clade in India, also has the highest frequency (19%) among the Austro-Asiatic tribals (Table 2.5). The coalescence time of the M2 haplogroup is been estimated (Kivisild et al. 1999b) to be $63000 \pm 6000$ ybp. (Because of small sample size of M2 in our data set, we have not estimated the coalescence time of this haplogroup.) Another subhaplogroup M4, whose frequency is about 15% in India (Figure 2.6), is completely absent among the Austro-Asiatic tribals. The estimated coalescence time of this haplogroup is $32000 \pm 7500$ ybp (Kivisild et al. 1999b). It is, therefore, likely that M4 arose after the expansion of the Austro-Asiatic tribals (estimated to have been about 55000 ybp; Table 2.6) and their entry into India. The M4 haplogroup either arose in India or was brought into India during a later wave of migration. Interestingly, the Austro-Asiatic tribals possess the minimum number of mtDNA RSP haplotypes (Table 2.2) and do not share many HVS1 sequences with other tribal groups (Figure 2.5), indicating that in spite of their antiquity they have remained essentially unacculturated, but exhibit the highest nucleotide diversity (Table 2.6), bolstering the view that they are the most

ancient tribal groups in India. We also note that there is clear indication of demographic expansion of all subgroups of tribals in India, but the expansion time of the Austro-Asiatic tribals is estimated to have been over 10000 years before those of the other subgroups. While we cannot be sure that this expansion took place within India, it is likely that the Austro-Asiatic tribals may have entered India prior to the other subgroups consequent to demographic expansion and its associated consequences of pressure on natural resources. In fact, Renfrew (1992) has contended that the initial dispersal process of modern humans out of Africa may have brought the Austric language group (to which Austro-Asiatic belongs) to approximately its present geographical distribution. Within India, the Austro-Asiatic speaking tribals are primarily concentrated in eastern and central India; and are absent from north or northwest India. However, one major tribal group (Khasi) of northeastern India also speak a dialect that belongs to the Austro-Asiatic subfamily. It, therefore, seems likely that the Austro-Asiatic speaking tribals may have entered India from the northeast, which is consistent with Diamond's (1997) view that this language subfamily evolved in southern China. It may also be noted that the Y-haplogroup 26, which is found in very high frequencies among the Han Chinese (Su et al. 2000; Tajima et al. 2001), is also found in high frequencies among the Austro-Asiatic tribals.

One clear finding from the analyses of separate sets of DNA markers is that the Tibeto-Burman speaking tribal populations, who are confined to the northeastern region of India, are clearly distinct from tribal populations belonging to the other language groups (Dravidian, Austro-Asiatic and Indo-European). Consistent with the anthropological view that the Tibeto-Burman tribals may have arrived in India in multiple waves and that that they may have experienced considerable admixture (Guha 1935), we have found that these groups possess the maximum number of mtDNA haplotypes, many of which are exclusive to them with low frequencies (Table 2.2). They also possess the lowest frequencies of the mitochondrial haplogroup U and the highest frequency of haplogroup D. Haplogroup D is known to be frequent in populations of Tibet and Korea , which reinforces the anthropological view (Guha 1935) that Tibetan, southern Chinese and south-east Asian populations are the ancestral populations of the Tibeto-Burmans tribals of India. Haplogroup U is, broadly speaking, a Caucasoid marker and the fact that Tibeto-Burman tribals of northeast India do not possess high frequencies of this haplogroup is consistent with the view (Thapar 1966) that the Indo-European speaking immigrants to India were not able to penetrate into this region of India. With respect to

the frequency distribution of Y-haplogroups, the Tibeto-Burman tribals are remarkably similar to the Austro-Asiatic tribals (Table 2.9 and Figure 2.13). Neither of these linguistic subgroups possesses the Y-haplogroup 3. This young haplogroup (estimated age is about 5500-7500 ybp; Table 2.12 and Karafet et al (1999)) is postulated (Zerjal et al. 1999) to be the most evident male legacy of the population expansion of the early nomadic groups of Central Asia. Therefore, it is not surprising that it is absent among the Austro-Asiatic and Tibeto-Burman tribals. These two groups of tribals also possess very high frequencies of the Y-haplogroup 26. However, the results of our discriminant analysis (Figure 2.15) show that the Tibeto-Burman and Austro-Asiatic tribal subgroups can be clearly differentiated with respect to the frequencies of 3 Y-STRP markers – DYS388, DYS391 and DYS393. Further, the results of our analyses based on autosomal markers also reveal that these two linguistic groups of tribals are genetically quite distinct (Figures 2.15 and 2.16). The Tibeto-Burman subfamily of the Sino-Tibetan language family has been subdivided (Grimes 1999) into four branches: Baric, Bodic, Burmese-Lolo and Karen. Based on a study of Y-chromosomal haplotypes, Su et al. (2000) have contended that after the proto-Tibeto-Burman people left their homeland in the Yellow River basin, the Baric branch moved southward and peopled the northeastern Indian region after crossing the Himalaya. This branch did not possess the YAP insertion element. Our findings are consistent with Su et al.'s (2000) inference.

Our result of the large extent of HVS1 sequence sharing between Indo-Europeans and Dravidians (Figure 2.5) appears surprising. There is overwhelming evidence from the distributions of mtDNA and Y-chromosomal haplogroups (Tables 3 and 5), and from the results of our analyses presented in Table 2.8 in which we have shown that the north Indian castes who speak Indo-European languages are significantly less differentiated from the Central Asian populations than population groups of the other regions of India who speak non-Indo-European languages, that there has been considerable gene flow from Central and West Asia into the northern regions of India. However, when we analyzed our data of Y-chromosomal and autosomal polymorphisms, it was revealed that the Dravidian tribal groups are similar to the Austro-Asiatic tribal groups, while the Dravidian caste groups are more similar to the Indo-European caste groups than to the Dravidian tribal groups (Table 2.9; Figures 2.14 and 2.18). The Y-haplogroup 9, which is commonly found in the Middle East (Hammer et al. 2000; Semino et al. 2000; Quintana-Murci 2001), has widely differing frequencies between Dravidian tribals

(9.5%) and castes (22.3%). The frequency among the Dravidian castes is closer to that of the Indo-Europeans. When these results are viewed along with the relative frequency distributions of the mtDNA haplogroups M and U (Figure 2.3), it appears that the Dravidian tribals and haplogroup M were spread throughout India prior to the Indo-European immigration into India. The Indo-European immigrants had brought with them a large influx of haplogroup U, although certain subhaplogroups of this haplogroup may have existed in India prior to the Indo-European immigration (Table 2.5 and Kivisild et al. 1999a). Thus, we find a decreasing gradient of haplogroup U frequency from north India to south India, and an inverse correlation of the haplogroup M and U frequencies over geographical space. It is also interesting that differences in frequencies of haplogroup U among linguistic groups is the more significant than among geographical regions or socio-cultural groups, as revealed by the logistic regression analysis. There is now clear evidence that the early social organization of the Indo-European immigrants was essentially tribal and that they were pastoral nomads or semi-nomads (Jha 1999). The Indo-Europeans later adopted agriculture (which was presumably brought into India from the Fertile Crescent region of West Asia) as a means of livelihood. The Dravidian speakers are said to have arrived in India from West Asia and brought with them the technology of agriculture. Settled life of the Indo-European semi-nomadic people led to the crystallization of a division of society and the formation of the ranked caste system. Many indigenous people of India embraced (or, were forced to embrace) the caste system, together with linguistic dominance and admixture. In fact, Renfrew (1992) has suggested that the elite dominance model, which envisages the intrusion of a relatively small but well-organized group that takes over an existing system by the use of force, may be appropriate to explain the distribution of the Indo-European languages in north India and Pakistan. As the Indo-Europeans, who entered India primarily through the northwest corridor about 3500 ybp, advanced into the Indo-Gangetic plain, indigenous people, especially the Dravidian speakers, who would not be linguistically dominated may have retarded southwards. Thus, after an initial period of admixture of the Dravidian people who adopted the caste system, they may have retarded to the southern region of India to avoid linguistic dominance. Thus, we find that there are considerable genetic similarities of the Indo-European castes with the Dravidian castes, although they presently occupy disjoint geographical territories. There is now clear evidence that the Harappan civilization, which flourished in northwestern region of the Indian subcontinent between 4500-3500 ybp, was not Indo-European (Kochhar

2000). A large fraction of the Dravidians who did not adopt the caste system, and therefore remained essentially unadmixed with the Indo-Europeans, comprise the Dravidian tribals of southern India. Hence, the Dravidian tribals show considerable genetic dissimilarities with the Dravidian castes. Surprisingly, although the Dravidians are said to have arrived in India from the Fertile Cresent region, with linguistic signatures of this movement remaining among the Elamites and Brahuis, the frequencies and patterns of Y-chromosomal haplogroups and STRP markers of the Dravidian tribals and castes of India are significantly dissimilar with those of the Brahuis, as a result of which the Brahuis do not cluster with the Dravidian tribals (Figure 2.14). This raises the question whether the existence of this Dravidian-speaking isolate in Pakistan can indeed be considered as a major indication of the Dravidian speakers arriving into India from West Asia through Pakistan. We also do not find that there is a smooth cline of Y-chromosomal HG-9 frequency from Iran through India (Mukherjee et al. 2001), as has been suggested by Quintana-Murci et al. (2001). As a matter of fact, the Dravidian-speaking tribals and castes possess quite dissimilar frequencies (9.5% and 22.3%, respectively) of this haplogroup. Another interesting finding of the present study is that the Indo-European speaking Halba tribals are most probably Dravidians who adopted an Indo-European language (Figure 2.17). Although we recognize that it is risky to make generalizations based on data of one group, we feel that the present Indo-European speaking tribals in India may not be the relics of any of the tribal populations who arrived in India from Central Asia, but may have been linguistically overpowered.

Our data also show that evidence that many contemporary tribal populations (e.g., Kota, Toda, Toto, Chakma) inhabiting disparate geographical regions and belonging to different linguistic groups have arisen by fission with a small number of founders. In such populations, therefore, find strikingly distinct genomic profiles. For example, the Kota and Toto possess only two of the 19-site HVS1 motifs. Similarly, several distinct motifs are found in high frequencies in a small number of populations (Figure 2.6). There are also large differences in haplotype diversities among populations (Figure 2.2 and other results not shown).

After the caste system was formed, a social practice that has evolved is that after marriage the wife moves to her husband's place of residence. This has obviously resulted in significantly greater movement of female, but not of male, lineages. This is discernible by a greater sharing of mtDNA haplotypes and HVS1 sequences, but not of Y-STRP haplotypes

(Table 2.11), a feature that has been noted earlier (Bamshad et al. 1998, Bhattacharyya et al. 1999). Consequently, we have found that the $F_{ST}$ values with respect to mtDNA markers are much smaller than those with respect to Y-chromosomal markers (Table 2.7).

We have observed that there is a fair concordance of ranks of $F_{ST}$ values based on mitochondrial, Y-chromosomal and autosomal markers (Table 2.7). As is observed, the $F_{ST}$ values based on autosomal markers are less than those for mitochondrial and Y-chromosomal markers, which is expected because the effective size of a population with respect to mt and Y markers is a quarter of that for autosomal markers. Hence, drift effects are expected to be more pronounced, resulting in a greater degree of genetic differentiation among populations, with respect to mt and Y markers compared to autosomal markers. The consistency of the results of our $F_{ST}$ and AMOVA analyses are very reassuring. The highest degree of differentiation is observed among tribal groups belonging to different geographical regions or linguistic groups. This is a testimony to their distinct ancestries, antiquities and isolation. The upper castes of different geographical regions are strongly differentiated with respect to mt and Y markers. However, for unclear reasons, they are not well-differentiated with respect to autosomal markers.

In summary, therefore, our results indicate that the tribal and the caste populations are strongly genetically differentiated. The Austro-Asiatic tribals are the earliest settlers in India and may have entered India through the northeast. They have remained essentially isolated for a long time. The significant extent of sharing of a small number of mtDNA haplotypes across populations indicates that the number of ancestral female lineages in India was small and also that there has been considerable female movement from one population to another. Most subsequent immigrations into India were predominantly of males. The Tibeto-Burmans tribals, who also entered India from the northeast, in spite of sharing genetic commonalities with the Austro-Asiatic tribals can be differentiated from them. The Dravidian tribals, who possibly entered India from the Fertile Crescent region, were once widespread throughout India. The Indo-European speaking nomads, who entered India through the northwest corridor from Central and West Asia, established their linguistic supremacy over a large number of Dravidian tribals and brought many of them under the fold of the Hindu caste system which they had formed after adopting a settled life. Many Dravidians, tribals and also castes, retreated to the southern regions possibly to retain their linguistic and other cultural identities from the Indo-European dominance.

The small number of tribal groups who currently speak Indo-European languages, were probably originally Dravidian speakers, but later adopted the Indo-European speech. Some cultural practices that evolved with the caste system have left their signatures on the genetic structures of contemporary populations. The upper castes of different geographical regions of India show a strong degree of genetic differentiation. Many contemporary populations were founded by small numbers of individuals, resulting in extremely deviant genetic profiles compared to populations that belong to the same geographical region or socio-cultural group. Historical migrations into India, resulting in differential admixtures with pre-existing population groups, have possibly contributed to a considerable obliteration of genetic histories of contemporary populations so that there is no clear congruence of genetic and geographical or socio-cultural affinities.

## Summary

We report a comprehensive study of a large number of ethnic populations of India based on mitochondrial, Y-chromosomal and autosomal markers. Our results indicate that the tribal and the caste populations are genetically highly differentiated. The four linguistic groups of tribals present in India, as also the upper castes of different geographical regions, are also highly differentiated. The Austro-Asiatic tribals seem to be the earliest settlers in India, as evidenced by their large nucleotide diversity and high frequencies of some ancient markers. Y-chromosomal haplogroup frequencies and their present geographical distribution indicate that may have entered India through the northeast. There is significant sharing of a small number of mtDNA haplotypes across populations indicating that the number of ancestral female lineages in India was small and also that there has been considerable female movement from one population to another. Subsequent immigrations into India appear to have been predominantly of males. The Tibeto-Burmans tribals, who also entered India from the northeast, share genetic commonalities with the Austro-Asiatic tribals but can be differentiated from them on the basis of Y-STRP haplotypes. The Dravidian tribals, who possibly entered India from the Fertile Crescent region, were possibly widespread throughout India, before the arrival of the Indo-European speaking nomads. After entering through the northwest Indian corridor from Central and West Asia. the Indo-Aryans established their linguistic supremacy over a large number of Dravidian tribals and brought many of them under the fold of the Hindu caste system, which they had formed after adopting a settled life. Many Dravidians, tribals and also castes, seem to have retreated to the southern regions possibly to retain their linguistic and other cultural identities. Indo-European trbals, were probably originally Dravidian speakers, but later adopted the Indo-European speech. Formation of populations by fission and cultural practices that evolved with the caste system have left their imprints on the genetic structures of contemporary populations. Historical migrations into India have contributed to a considerable obliteration of genetic histories of contemporary populations so that there is currently no clear congruence of genetic and geographical or socio-cultural affinities.

## Chapter 2: References


Ballinger SW, Schurr TG, Torroni A, Gan YY, Hodge JA, Hassan K, Chen K-H, Wallace DC (1992) Southeast Asian mitochondrial DNA analysis reveals continuity of ancient mongoloid migrations. Genetics 130: 139-152

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BVR, Rasanayagam A, Hammer MF (1998) Female gene flow stratifies Hindu castes. Nature 395: 851-852

Bamshad MJ, Kivisild T, Watkins WS, Dixon MP, Ricker LE, Rao BB, Naidu M, Prasad BVR, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of Indian caste populations. Genome Res 11: 994-1004

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37-48

Beteille A (1998) The Indian heritage – a sociological perspective. In: Balasubramian D, Appaji Rao N (eds) The Indian human heritage. University Press, Hyderabad. pp 87-94

Bhattacharya A (1946) On a measure of divergence between two multinomial populations. Sankhya 7: 401-406

Bhattacharyya N, Basu P, Das M, Pramanik S, Banerjee R, Roy B, Roychoudhury S, Majumder PP (1999) Negligible gene-flow across ethnic boundaries in India, revealed by analysis of Y- chromosomal DNA polymorphisms. Genome Res 9: 711-719

Bonatto SL, Salzano FM (1997) A single and early origin for the peopling of the Americas supported by mitochondrial DNA sequence data. Proc Natl Acad Sci USA 94: 1866-1871

Buxton LHD (1925) The peoples of Asia. Hutchinson, London

Calafell F, Underhill P, Tolun A, Angelicheva D, Kalaydjieva L (1996) From Asia to Europe: mitochondrial DNA sequence variability in Bulgarians and Turks. Ann Hum Genet 60 :35-49

Cann RL (2001) Genetic clues to dispersal of human populations: Retracing the past from the present. Science 291: 1742-1748

Casanova M, Leroy P, Boucekkine C, Weissenbach J, Bishop C, Fellous M, Purrello M, Fiori G, Siniscalco M (1985) A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science 230: 1403-1406

Cavalli-Sforza LL (2000) Genes, peoples, and languages. University of California Press, Berkeley

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

Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Martinez-Arias R, Clarimon J, et al (1998) Trading genes along the silk road: mtDNA sequences and the origin of Central Asian populations. Am J Hum Genet 63: 1824-1838

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet 70: 1197-1214

Diamond J (1997) Guns, germs and steel: the fates of human societies. Jonathan Cape, London

Disotell TR (1999) Human evolution: the southern route to Asia. Curr Biol 9:R925-R928

Endicott P, Thomas M, Gilbert P, Stringer C, Lalueza-Fox C, Willerslev E, Hansen AJ, Cooper A (2002) The genetic origins of the Andaman Islanders. Am J Hum Genet (in press)

Forster P, Torroni A, Renfrew C, Rohl A (2001) Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. Mol Biol Evol 18: 1864-1881

Fu Y-X (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-925

Gadgil M, Joshi NV, Shambuprasad UV, Manoharan S, Patil S (1998) Peopling of India. Balasubramian D, Appaji Rao N (eds) The Indian human heritage. University Press, Hyderabad. pp 100-129

Grimes BF (1999) The ethnologue: languages of the world. Summer Institute of Linguistics, California

Guha BS (1935) The racial affinities of the people of India. in Census of India, 1931, Part III - Ethnographical. Government of India Press, Simla

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol 15: 427-441

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T, Ostrer H, Bonne-Tamir B (2000) Jewish and middle eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci USA 97: 6769-6774

Harpending HC, Sherry ST, Rogers AR, Stoneking M (1993) The genetic structure of ancient human populations. Curr Anthrop 34: 483-496

Hudson RR, Slatkin M, Maddison WP (1992) Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589

Jha DN (1999) Ancient India in historical outline. Manohar Publishers, New Delhi

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, Soodyall H, Jenkins T, Rogers AR (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. 57: 523-538

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergan A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Weibe V, Griffiths RC, Templeton AR, Hammer MF (1999) Asian source(s) of New World Y-chromosome founder haplotypes. Am J Hum Genet 64: 817-831

Karve I (1961) Hindu society: an interepretation. Deshmukh Prakashan, Poona.

Keith A (1936) Review of B.S. Guha's Racial Affinities of the Peoples of India. Man 29: 37

Kennedy KAR, Deraniyagala SU, Roertgen WJ, Chiment J and Sisotell T (1987) Upper Pleistocene fossil hominids from Sri Lanka. Amer J Phys Anthrop 72: 441-461

Kidd KK (1998) A global survey of haplotype frequencies and linkage disquilibrium at the DRD2 locus. Hum Genet 103: 211-227

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, Parik J, Watkins WS, Dixon ME, Papiha SS, Mastana SS, Mir MR, Ferak V, Villems R (1999a) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9: 1331-1334

Kivisild R, Kaldma K, Metspalu M, Parik J, Papiha S, Villems R (1999b) The place of the Indian mitochondrial DNA variants in the global network of the maternal lineages and the peopling of the old world. In: Papiha SS, Deka R, Chakraborty R (eds) Genome diversity: applications in human population genetics. Kluwer, New York. Pp. 135-152

Kochhar R (2000) The vedic people: their history and geography. Orient Longman, New Delhi

Kosambi DD (1991) The culture and civilisation of ancient India in historical outline. Vikas Publishing House, New Delhi

Kulozik AE, Wainscoat JS, Serjeant GR, Kar BC, Al-Awamy B, Essan GJF, Falusi AG, Haque SK, Hilali AM, Kate SL, Ranasinghe WAEP, Weatherall DJ (1986) Geographical survey of HbS gene haplotypes: evidence for an independent Asian origin of the sickle-cell mutation. Am J Hum Genet 39: 239-244

Lahr MM, Foley RA (1994) Multiple dispersals and modern human origins. Evol Anthrop 3: 48-60

Lahr MM, Foley RA (1998) Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. Yrbook Phys Anthrop 41: 137-176

Lundstrom RS, Tavare S, Ward RH (1992) Estimating substitution rates from molecular data using the coalescent. Proc Natl Acad Sci USA 89: 5961-5965

Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2: 13-20

Majumdar P and Majumder PP (2000) HAPLOFREQ: a computer program package for maximum-likelihood estimation of haplotype frequencies in a population via the EM algorithm. Tech Rep AHGU-1/2000, Indian Statistical Institute, Calcutta.

Majumder PP (1998) People of India: Biological diversity and affinities. Evol Anthrop 6: 100-110

Majumder PP, Roy B, Banerjee S, Chakraborty M, Dey B, Mukherjee N, Roy M, Guha Thakurta P, Sil SK (1999a) Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. Eur. J. Hum. Genet. 7: 435-446

Majumder PP, Roy B, Balgir RS, Dash BP (1999b) Polymorphisms in the beta-globin gene cluster in some ethnic populations of India and their implications on disease. In: Gupta S, Sood OP (eds) Molecular intervention in disease. New Delhi: Ranbaxy Science Foundation. pp. 75-83

Meenakshi K (1995) Linguistics and the study of early Indian history. In: Thapar R (ed) Recent perspectives of early Indian history. Popular Prakashan, Bombay. pp 53-79

Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16, 1215

Misra VN (1992) Stone age in India: an ecological perspective. Man and Env. 14:17-64

Mohiyuddin A, Ayub Q, Qamar R, Zerjal T, Helgason A, Mehdi SQ, Tyler-Smith C (2001) Y-chromosomal STR haplotypes in Pakistani populations. Forensic Sci Intl 118: 141-146

Mukherjee N, Nebel A, Oppenheim A, Majumder PP (2001) High-resolution analysis of Y-chromosomal polymorphisms reveals signatures of population movements from Central Asia and West Asia into India. J Genet 80: 125-135

Nebel A, Filon D, Weiss DA Weale M, Faerman M, Oppenheim A, Thomas MG (2000) High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews. Hum Genet 107: 630-641

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321-3323

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

Nei M, Jin L (1989) Variances of the average number of nucleotide substitutions within and between populations. Mol Biol Evol 6: 290-300.

Nei M, Ota T (1991) Evolutionary relationships of human populations at the molecular level. In: Osawa S, Honjo T (eds) Evolution of life. Springer, Tokyo. pp. 415-428

Parpola A (1975) On the protohistory of the Indian languages in the light of archaeological, linguistic and religious evidence: an attempt at integration. in *South Asian Archaeology* (ed. van Lohuizen-De Leeuw, J.E.) pp. 73-84. Brill Academic Publishers, New York.

Passarino G, Semino O, Modiano G, Bernini LF, Santachiara-Benerecetti AS (1996a) mtDNA provides the first known marker distinguishing proto-Indian from the other Caucasoids; it probably predates the diversification between Indians and Orientals. Ann Hum Biol 23: 121-126

Passarino G, Semino O, Bernini LF, Santachiara-Benerecetti AS (1996b) Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. Am J Hum Genet 59: 927-934

Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M, Santachiara-Benerecetti AS (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. Am J Hum Genet. 62:420-34

Pattanayak DP (1998) The language heritage of India. in The Indian human heritage (eds. Balasubramanian, D., Rao, N.A.) pp. 95-99. Universities Press, Hyderabad

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. Am J Hum Genet 65:208-219

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959

Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ (2002) Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107-24

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. Nat Genet 23: 437-441

Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K (2001) Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. Am J Hum Genet 68: 537-542.

Rapson EJ (1955) Peoples and languages. In Rapson EJ (ed.) *Cambridge History of India, Vol. 1: Ancient India.* S. Chand & Co., Delhi

Ratnagar S (1995) Archaeological perspectives of early Indian societies. In: Thapar R (ed) Recent perspectives of early Indian history. Popular Prakashan, Bombay. pp 1-52

Ray N (1973) Nationalism in India. Aligarh Muslim University, Aligarh

Redd A, Roberts-Thomson J, Karafet T, Bamshad M, Jorde LB, Naidu JM, Walsh B, Hammer MF (2002) Gene flow from the Indian subcontinent to Australia: evidence from the Y-chromosome. Curr Biol 12: 673-677

Renfrew C (1987) Archaeology and the puzzle of Indo-European origins. Jonathan Cape, London

Renfrew C (1992) Archaeology, genetics and linguistic diversity. Man 27: 445-478

Risley HH (1915) The people of India. Thacker Spink, Calcutta

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper G, Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previdere C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA (2000) Y-chromosomal diversity in Europe is clinal and is influenced primarily by geography, rather than by language. Am J Hum Genet 67: 1526-1543

Roychoudhury S, Roy S, Basu A, Banerjee R, Vishwanathan H, Usha Rani MV, Sil SK, Mitra M, Majumder PP (2001) Genomic structures and population histories of linguistically distinct tribal groups of India. Hum Genet 109: 339-350

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15: 174-175

Ruhlen M (1991) *A Guide to the World's Languages.* Stanford University Press, Stanford

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425

Sarkar SS (1958) Race and race movements in India. In *The Cultural Heritage of India*, Vol. 1 (ed. Chatterjee, S.K.). pp. 17-32. The Ramakrishna Mission Institute of Culture, Calcutta

Schneider S, Kueffer J-M, Roessli D, Excoffier L (2000) ARLEQUIN: A software for population genetic data analysis. University of Geneva, Geneva

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, de Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benercetti AS, Cavalli-Sforza LL, Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y-chromosome perspective. Science 290: 1155-1159

Singh KS (1992) People of India: an introduction. Anthropological Survey of India, Calcutta

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129: 555-562

Stringer C (2000) Coasting out of Africa. Nature 405: 24-25

Su B, Xiao C, Deka R, Seielstad MT, Kangwanpong D, Xiao J, Lu D, Underhill P, Cavalli-Sforza LL, Chakraborty R, Jin L (2000) Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum Genet 107: 582-590

Tajima A, Pan I –H, Fucharoen G, Fucharoen S, Matsuo M, Tokunaga K, Juji T, Hayami M, Omoto K, Horai S (2001) Three major lineages of Asian Y chromosomes: implications for the peopling of east and southeast Asia. Hum Genet 110: 80-88

Thapar R (1966) A history of India, volume 1. Penguin, Middlesex

Thapar R (1995) The first millennium B.C. in northern India (upto the end of Mauryan period). In: Thapar R (ed) Recent perspectives of early Indian history. Popular Prakashan, Bombay. pp 80-141

Thomas M, Bradman N, Flinn H (1999) High throughput analysis of 10 microsattelite and 11 diallelic polymorphisms on the human Y-chromosome. Hum.Genet 105:577-581

Tishkoff SA, Ruano G, Kidd JR, Kidd KK (1996) Distribution and frequency of a polymorphic *Alu* insertion at the plasminogen activator locus in humans. Hum. Genet. 97: 759-764

Torroni A, Schurr TB, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53: 563-590

Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, Wallace DC (1994) mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am J Hum Genet 55: 760-76

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, Savontaus M-L, Wallace DC (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144: 1835-1850

Vigilant LA, Wilson AC, Harpending H (1991) African populations and the evolution of human mitochondrial DNA. Science 253, 1503-1507

Wallace DC (1995) Mitochondrial DNA variation in human evolution, degenerative disease and aging. Am J Hum Genet 57: 201-223

Watkins WS, Bamshad M, Dixon ME, Bhaskara Rao B, Naidu JM, Reddy PG, Prasad BV, Das PK, Reddy PC, Gai PB, Bhanu A, Kusuma YS, Lum JK, Fischer P, Jorde LB (1999) Multiple origins of the mtDNA 9-bp deletion in populations of South India.Am J Phys Anthropol 109:147-158

Wells RS, Yuldasheva N, Ruzibakiev R, Underhill PA, Evseeva I, Blue-Smith J et al. (2001) The Eurasian heartland: A continental perspective on Y-chromosome diversity. Proc Natl Acad Sci (USA) 98: 10244-10249

Zerjal T, Pandya A, Santos FR, Adhikari R, Tarazona-Santos E, Kayser M, Evgrafov OV, Singh L, Thangaraj K, Destro-Bisol G, Thomas M, Qamar R, Mehdi SQ, Rosser ZH, Hurles ME, Jobling MA, Harihara S, Tyler-Smith C (1999) The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. In: Papiha SS, Deka R, Chakraborty R (eds) Genomic diversity: applications in human population genetics. Plenum, New York, pp 91-101

Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C (2002) A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. Am J Hum Genet 71: 466-482

# CHAPTER 3

## Identification of Polymorphic Motifs Using Probabilistic Search Algorithms

## Introduction

Single nucleotide polymorphisms (SNPs) that occur in the human geneome at roughly 1 per 2 kb spacing on the average (Balasubramanian et al. 2002) are often phylogenetically associated. Various evolutionary mechanisms, including natural selection, maintain the association of specific variant nucleotides at one or more sites, which may not be contiguous. The search for associated nucleotides at a set of polymorphic positions is of interest in studies of common diseases (Sabeti et al. 2002) and in evolutionary genetics (Tateno et al. 1997; Daly et al. 2001). We define a set of nucleotides that occurs at a high frequency at multiple polymorphic DNA sites, not necessarily contiguous, in a group of individuals as a "motif". We note that our definition of a motif differs from the conventional definition, as for example that is used for finding regulatory sequences in promoter regions of genes (Keiler and Shapiro 2001), in two ways: (a) the sites included in our motif definition are polymorphic and (b) the sites may not be contiguous. In conventional problems, search is made for evolutionary conserved motifs at a contiguous set of nucleotide positions (Gupta and Liu 2003). In case-control studies of common diseases, it is of interest to find such motifs and to test whether there are differences in motif frequencies between cases and controls (Khani-Hanjani et al. 2002). Motifs that are found in significantly higher frequencies among cases are associated with the disease under study. If variants in multiple genes are indeed involved in the disease, the sites in such a motif may not be contiguous. Similarly, the discovery of such motifs is important in evolutionary genetics. Indeed such motifs have been used to define subhaplogroups of specific clades (haplogroups) of the human mitochondrial (mt) DNA (Bamshad et al. 2001).

It is theoretically possible to discover polymorphic motifs in a set of $N$ aligned DNA sequences, each of length $L$ nucleotides, by examining frequencies in all possible $k \times k$ tables, $k$=2,3,...,$L$. However, this is computationally infeasible. The purpose of this chapter is to propose a set of probabilistic search algorithms that may be used for motif finding under different scenarios, and to evaluate their efficiencies using both synthetic and real data sets.

## Definition of the Problem

Consider a data matrix $((a_{ij}))_{N\times L}$, where $a_{ij}$ denotes the nucleotide (A,T,G or C) at the j-th polymorphic site (j=1,2,...,L) for the i-th individual (i=1,2,...,N). The data matrix is generated from aligned DNA sequences of a specific genomic segment of N individuals, from which all monomorphic sites have been removed. We note that if these N individuals belong to a case-control study, then the data matrix needs to be initially created by pooling all cases and controls, and subsequently separated into two matrices one for case and the other for controls. Similar caution is also required in evolutionary studies while simultaneously dealing with two populations. We also note that if disjoint segments of DNA are to be simultaneously examined for motif finding, then appropriate segments may be separately aligned and the aligned segments concatenated in the data matrix.

Let V={1,2,...,L} denote the set of all L polymorphic sites in the data. Let $\Pi_p$ denote the set of all possible combinations of p sites from V. To fix ideas, consider the data matrix given in Table 3.1. In this matrix, N=4, and L=7. Here, V={1,2,...,7}. For p=2, $\Pi_2$ ={{1,2},{1,3},...,{1,7},{2,3},{2,4},...,{6,7}}=$\{V_2^k\}$, k=1,2,...,$\binom{7}{2}$. In general, $\Pi_p = \{V_p^k\}$, k=1,2,...,$\binom{L}{p}$ and $V_p^k = \{x_1^k, x_2^k, \ldots, x_p^k : x_i^k \in V\}$. We define the modal sequence on $V_p^k$ as that particular combination of nucleotides at the sites $\{x_1^k, x_2^k, \ldots, x_p^k\}$ included in $V_p^k$, k=1,2,..., $\binom{L}{p}$. In the data matrix of Table 3.1, the modal sequence, for example, on $V_2^1$={1,2\} is AG with frequency 2, $V_2^2$={1,3} is AT with frequency 3, etc. We define a motif of length p as the maximally frequent modal sequence on $\Pi_p$; that is, the sequence that occurs with the highest frequency (globally modal) among modal sequences on $V_p^1, V_p^2, \ldots, V_p^{\binom{L}{p}}$. In our example, the motif of length 2 is AT on $V_2^2$={1,3} with frequency 3. In general, the problem of finding a motif of length p from a N×L data matrix reduces to identifying the set $V_p^k$, k=1,2,..., $\binom{L}{p}$, from $\Pi_p$, such that the modal sequence on $V_p^k$ is globally modal. With a N×L data

**Table 3.1**
**An Example of a Data Matrix**

| Sequence/Individual No. | Variant Site No. | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | A | A | T | T | G | C | C |
| 2 | A | G | T | C | G | C | T |
| 3 | A | G | T | T | A | C | T |
| 4 | G | G | C | C | A | T | T |

matrix, the search space $\Pi_p$ has $\binom{L}{p}$ elements. Hence, an exhaustive search of this space is computationally very expensive, and perhaps infeasible. We propose a stochastic search method, similar in spirit to Metropolis-Hastings version of simulated annealing. Our objective function, E(S), to be maximized is the "frequency of a string (S) of nucleotides at p out of L sites".

## Probabilistic Search Algorithms for a Given Value of Motif Length

Instead of maximizing E(S), we shall consider minimizing a monotonically decreasing function, H(S), of E(S). The algorithm is iterative. We start with a string $S_0$ of length p; that is, a set of p distinct nucleotide sites drawn randomly from the L polymorphic sites. In each iterative step an element (a nucleotide at a specific site) of the string $S_0$ is updated. Hence, after p such steps we get a completely updated string. The procedure of updating $S_0$ to $S_1$ is called a *sweep*. Thus, *a sweep comprises p iterative steps*. Let $S_t$ denote the updated string after t sweeps.

We shall use the following notations:

1. Let $S_t = (x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(p)})$.
2. Let $S_t^{(i)}$ denote a string in the (t+1)-th sweep, whose first i ($0 \leq i \leq p-1$) sites have already been modified.
3. Let $S_t^{(i)}(y)$ denote a string in the (t+1)-th iteration, whose first i ($0 \leq i \leq p-1$) sites have already been modified and the (i+1)-th site is replaced by site y.
4. Let $H_t^{(i)}$ denote the minimum value that has been found for H(S) in the course of all the necessary evaluations of H(S) till the completion of the i-th iterative step in the (t+1)-th sweep.
5. Let $M_t^{(i)}$ denote the string corresponding to $H_t^{(i)}$.

We initially set $H_0^{(0)}=0$ and $M_0^{(0)}$ as a "null" string. The updating procedure for the i-th element in the (t+1)-th sweep uses the idea underlying the Metropolis-Hastings algorithm (Metropolis et al. 1953) which can be described as follows:

We first calculate $\beta_t = c.\ ln(t+1)$; where c is a constant > 0. One site (x) is selected at random from the set $V \backslash S_t^{(i-1)}$; that is, from the set V={1,2,...,L} from which the sites included in the set $S_t^{(i-1)}$ have been removed. We then probabilistically update $x_t^{(i)}$ to $x_{t+1}^{(i)}$ according the following rule:

$$x_{t+1}^{(i)} = \begin{cases} x & \text{with probability } \min(\Lambda,1) \\ x_t^{(i)} & \text{with probability } 1 - \min(\Lambda,1) \end{cases}$$

where, $\Lambda = \exp[-\beta_t \{H(S_t^{(i-1)}(x)) - H(S_t^{(i-1)}(x_t^{(i)}))\}]$.

Obviously, the transition probability from one string to another depends only on the outcome of the previous transition (Markov Property). As is easily understood from the above updating rule, at any step of the iteration, although a string that yields a smaller value of H(S) is accepted with a high probability, to avoid being trapped at a local minimum, the current string with higher value of H(S) may also be retained with a small probability (that crucially depends on the control parameter c).

After each iteration we compare $H(S_t^{(i-1)}(x))$ with $H_t^{(i-1)}$. If, $H(S_t^{(i-1)}(x)) < H_t^{(i-1)}$ then, $H(S_t^{(i-1)}(x))$ is the new value for $H_t^{(i-1)}$ and $M_t^{(i)}$ is the updated string $S_t^{(i-1)}(x)$. Otherwise, we do not change $H_t^{(i-1)}$ and $M_t^{(i)}$. In each iteration, therefore we compare the value of the objective function with the smallest value it has attained thus far. This introduces the concept of elitism, which is popular in evolutionary computation (Bhandari et al. 1996), in our algorithm and is done to avoid being trapped at a local minimum.

The possible stopping rules for terminating sweeps in our algorithm can be:

(i) stop if an upper bound, usually a large preset number dependent on availability of computing resources, on the total number of sweeps (including new initials, if any) is reached, and

(ii) check if the minimum value of H(S) attained thus far during the algorithm has remained unchanged for a certain (preset) number of sweeps. If so, terminate.

We note that as with all numerical optimization procedures, it is desirable to repeat the procedure a certain number of times from different initial starting values, in the present procedure from different initial strings, and examine whether convergence to the same optimal value is obtained. The number of repetitions of the procedure that is practically feasible obviously depends on the availability of computing resources.

## Performance of the Algorithm with a Given Motif Length: Assessment Using Synthetic Data Sets

*Data Set 1*: We designed various synthetic data sets, so that the motif in each data set was known, to assess the performance of our algorithm. In the synthetic data matrix - I of size 1000×50, comprising data on 1000 (=N) individuals at 50 (=L) binary polymorphic sites, variants at the first 10 sites were assigned non-randomly, while variants at the remaining 40 sites were randomly assigned. This synthetic data structure (Table 3.2) permitted us to assess convergence of our algorithm for different motif lengths; the expected frequency, E(S), of the motif of maximum length 10 [$S = ((1))_{1\times10}$ = (1,1,1,1,1,1,1,1,1,1)] at the first 10 sites, that is $x_i = i$ (i = 1,2,...,10), being 750. We carried out 1000 independent runs with p = 10 (given motif length). In each run, the "random submatrix" of the data matrix was generated. We have taken $H(S) = \frac{1}{1+E(S)}$, in this and in all the remaining analyses, unless otherwise mentioned. The results are presented in Table 3.3 from which it is seen that our algorithm identified the motif with a success rate of 100%. The number of sweeps to convergence varied with different initial values of the control parameter, c. Fastest convergence was achieved for c = 200.

*Data Set 2*: While the first set synthetic data was, in some sense, the best-case scenario since the motif could be easily discovered, we constructed the second synthetic data set in which the motif, although easily identified, is hard to discover. The structure of this second data matrix is given in Table 3.4. For a given motif length p=10, it is clear that the motif is $S = ((1))_{1\times10}$ at the sites 41-50, with E(S) = 900. However, this motif is clearly

Table 3.2
Structure of the Synthetic Data Matrix-1

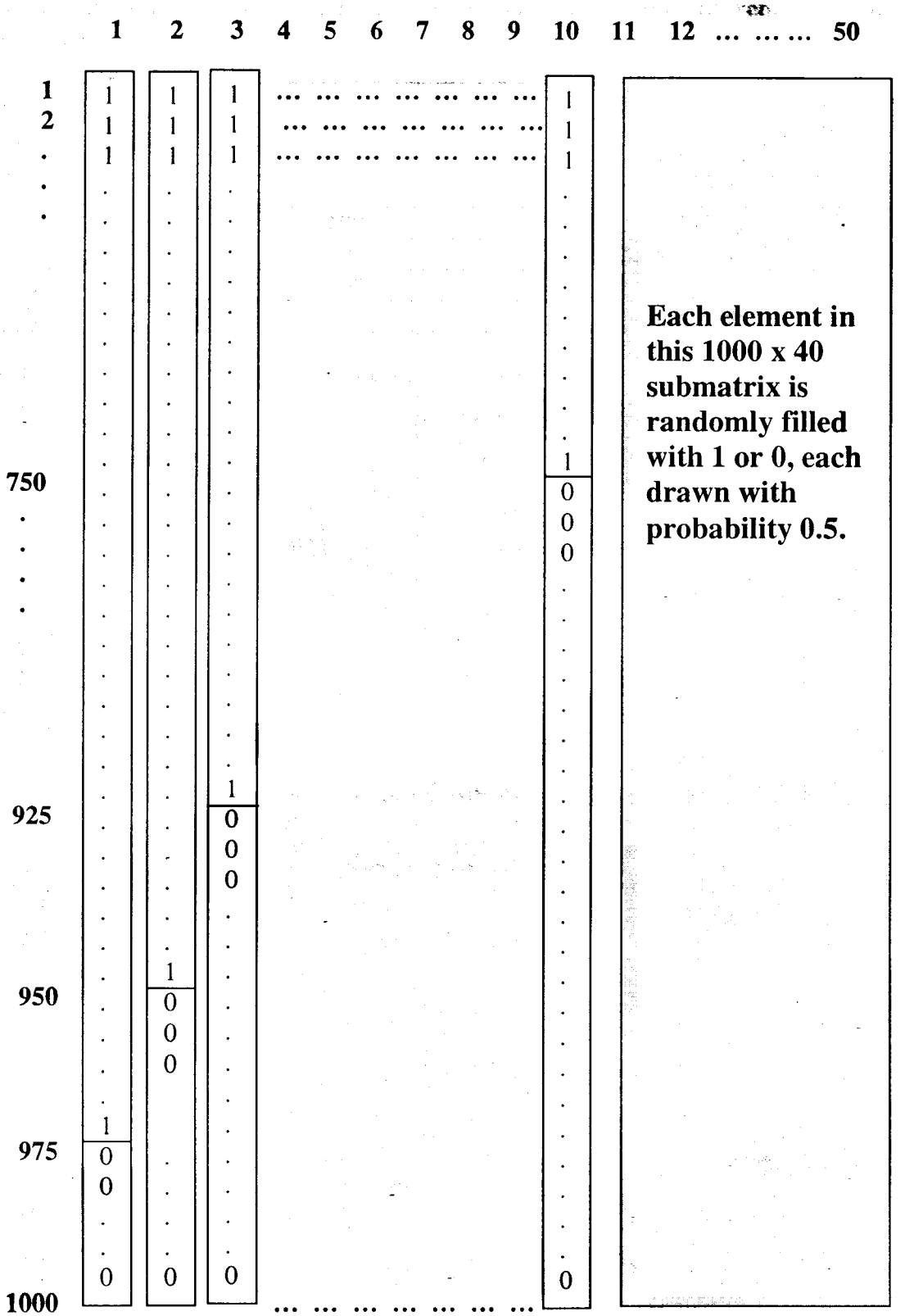
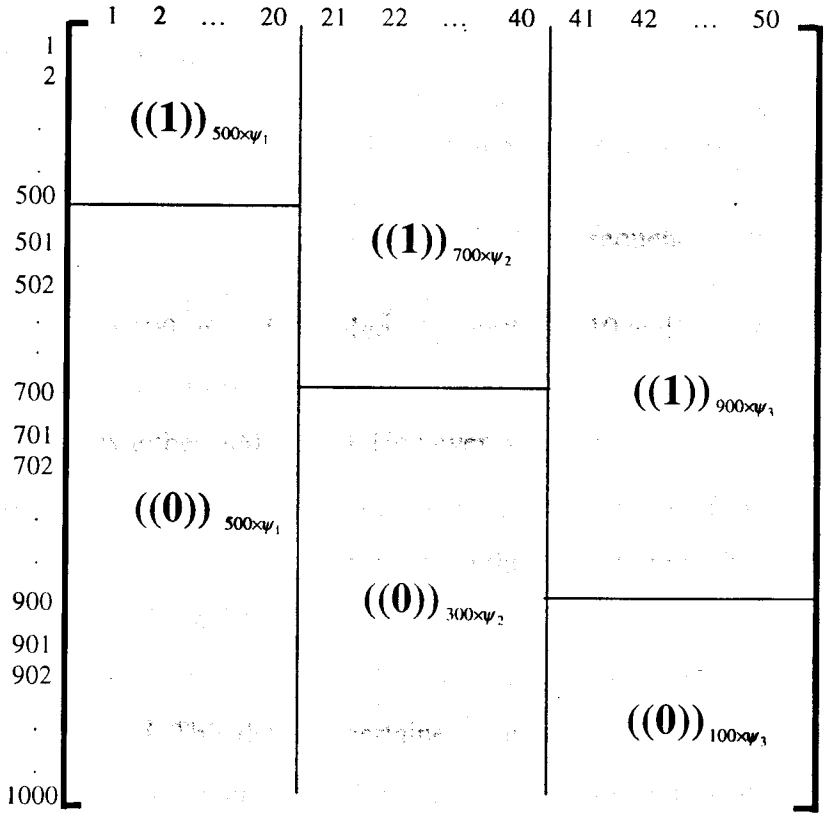


**Table 3.3**
**Results of 1000 Simulation Runs for Synthetic Data Set 1, with Different Values of the Control Parameter *c***

| Initial value of *c* | % of runs in which the "correct" motif was identified | Mean ± s.d. of the number of sweeps to convergence |
|---|---|---|
| 10 | 100 | 638.47 ± 283.3 |
| 30 | 100 | 93.73 ± 44.8 |
| 50 | 100 | 72.48 ± 42.3 |
| 100 | 100 | 64.06 ± 42.7 |
| 200 | 100 | 55.51 ± 33.8 |
| 250 | 100 | 60.21 ± 37.9 |

**Table 3.4**
**Structure of the Synthetic Data Matrix – 2**




$\psi_1$ = No. of columns in the first set of submatrices = 20 [comprising columns (1,2,…,20)],

$\psi_2$ = No. of columns in the second set of submatrices = 20 [comprising columns (21,22,…,40)],

$\psi_3$ = No. of columns in the third set of submatrices = 10 [comprising columns (41,42,…,50)].

almost impossible to find, because there is exactly one such among the $\binom{50}{10}$ possibilities. Among these $\binom{50}{10}$ points in the search space, there are $\left[\binom{50}{10}-\binom{30}{10}\right]$ 10-site combinations at which the sequence will be $((1))_{1\times10}$ with a frequency of 500, and $\left[\binom{30}{10}-1\right]$ 10-site combinations at which the sequence will be $((1))_{1\times10}$ with a frequency of 700. In 100 runs of our algorithm with p = 10 and different initial values of c, we were never able to discover the correct motif. Invariably, the convergence was to a string with frequency either 500 or 700. However, when $(\psi_1,\psi_2,\psi_3)$ was changed from (20,20,10) to other combinations of values keeping the structure of the data matrix similar (see Table 3.4), the proportions of runs in which the correct motif of length 10 was discovered increased (Table 3.5).

*Data Set 3*: This data set pertained to evolution. When two populations that have diverged from an ancestral population evolve separately, it is often found that the daughter populations accumulate separate sets of mutations that increase in frequencies because of natural selection. Thus, one finds motifs in the daughter populations, with some motif sites that are shared between the two populations, while some are unshared. (Schwaiger and Epplen 1995). The shared sites are presumably those sites that belonged to a motif which was present in the ancestral population, while the unshared sites are those that have arisen and increased in frequency since the divergence of the two populations from the ancestral population. We constructed a synthetic data set to mimic this evolutionary scenario and applied our algorithm to assess whether it is possible to discover the relevant motifs. For constructing the data set, we first created a data matrix of size 1000×50, and assigned a value of 0 or 1 to each cell with probability 0.5. Then, we randomly selected 10 sites (columns of the data matrix) from the set ($\Pi$) of 50 sites, and changed the 0s to 1s at each site (column), so that at each of these sites the proportion of 1s among the 1000 individuals was $\geq 0.8$. This resulted in the data matrix, $D_1$, of the ancestral population which expectedly has a motif of length 10 comprising the set of 10

**Table 3.5**
**Results of 1000 Simulation Runs for Different Stuctures of Synthetic Data Set 2 as Specified by $(\psi_1,\psi_2,\psi_3)$[1] with $c$=200**

| $(\psi_1,\psi_2,\psi_3)$ | % of runs in which the "correct" motif was identified | Mean ± s.d. of the number of sweeps to convergence |
|---|---|---|
| (20,20,10) | 0 | -- |
| (19,19,12) | 1.60 | 1372.25 ± 497.7 |
| (18,18,14) | 15.2 | 1082.18 ± 542.7 |
| (17,17,16) | 55.0 | 896.77 ± 571.4 |

[1] See Table 3.4 for definitions of $(\psi_1,\psi_2,\psi_3)$

randomly selected sites, which we shall denote as $\Pi_1$. We then created two daughter populations of this ancestral population. The data matrix, of size $1000 \times 50$, corresponding to the first daughter population was initially created by sampling 1000 rows (each with 50 columns), with replacement, from the data matrix of the ancestral population. We then selected a set ($\Pi_2$) of 5 sites randomly from $\Pi \backslash \Pi_1$, and at these selected sites we randomly replaced 0s by 1s in the initial data matrix of the first daughter population such that the proportions of 1s among the 1000 individuals at each of these 5 sites was $\geq 0.8$. This yielded the final data matrix, $D_2$, corresponding to the first daughter population in which the motif expectedly comprises sites belonging to $\Pi_1 \cup \Pi_2$ of length 15. For the second daughter population, the initial data matrix was similarly created. A set ($\Pi_3$) of 5 sites were chosen from $\Pi \backslash \Pi_1 \cup \Pi_2$ and the final data matrix, $D_3$, was similarly created. In the second daughter population, the expected motif of length 15 has sites belonging to $\Pi_1 \cup \Pi_3$.

We carried out 1000 independent simulation runs using the procedure described above, with c=200. Detailed results for 5 runs are provided in Table 3.6, which shows that our probabilistic search algorithm always converged to the "correct" motifs of "correct" lengths in a small number of sweeps. As a matter of fact, correct convergence was achieved in each of the 1000 runs (detailed results not provided). The mean $\pm$ s.d. of the number of sweeps to convergence was obviously different for the parental and the daughter populations. These values are provided in Table 3.7, which show that convergence using the proposed algorithm is fairly fast.

## Refinement of the Probability Search Algorithm when Motif Length is Unknown

In reality, the motif length (p) is usually unknown. Of course, when p is unknown, it is possible to start with a small value of p and increase this value sequentially, examining for each value of p the extent of decrease of the value of E(S). One can stop with that value of p, when an increase to (p+1) results is a "substantial" drop in the value

**Table 3.6**

**Detailed Results Pertaining to Synthetic *Data Set-3* for 5 Independent Simulation Runs, with Stopping Rule "Stop if Number of Sweeps = 2000 "**

| Characteristics of Synthetic Data Matrices | | | | | Number of Sweeps to Convergence | Whether Converged to "Correct" Motif |
|---|---|---|---|---|---|---|
| Population (Data Matrix) | Motif length | Simulation Number | Sites in Motif[1] | Frequencies of "1" at Motif Sites | | |
| Population 1 ($D_1$) | 10 | 1 | 17 22 24 27 28 36 39 40 43 50 | 855 845 847 837 855 847 843 847 827 841 | 38 | YES |
| | | 2 | 1 4 8 15 19 20 22 37 39 44 | 855 832 848 867 861 874 850 858 851 849 | 27 | YES |
| | | 3 | 6 7 8 9 11 16 20 21 35 47 | 876 846 837 852 830 855 843 851 859 849 | 64 | YES |
| | | 4 | 3 11 13 27 35 39 42 46 47 50 | 846 868 851 834 874 874 837 848 832 856 | 35 | YES |
| | | 5 | 5 7 10 16 22 29 30 33 42 44 | 830 846 827 840 851 860 853 851 858 826 | 84 | YES |
| Population 2 ($D_2$) | 15 | 1 | 17 22 24 27 28 36 39 40 43 50 ***18 31 34 42 48*** | 863 845 867 854 875 846 851 858 844 832 845 827 816 888 855 | 13 | YES |
| | | 2 | 1 4 8 15 19 20 22 37 39 44 ***7 12 21 31 46*** | 857 821 852 882 863 887 838 866 858 853 828 890 886 830 850 | 37 | YES |
| | | 3 | 6 7 8 9 11 16 20 21 35 47 ***2 27 36 37 41*** | 873 835 839 859 828 853 863 851 876 848 845 846 890 824 816 | 77 | YES |
| | | 4 | 3 11 13 27 35 39 42 46 47 50 ***5 15 26 37 38*** | 844 874 848 831 867 860 819 862 834 856 827 809 886 826 856 | 11 | YES |
| | | 5 | 5 7 10 16 22 29 30 33 42 44 ***1 8 27 35 46*** | 843 840 821 848 854 880 838 836 847 820 821 845 876 860 881 | 28 | YES |
| Population 3 ($D_3$) | 15 | 1 | 17 22 24 27 28 36 39 40 43 50 ***3 9 15 33 44*** | 870 843 847 841 824 857 859 864 822 845 882 828 811 892 852 | 42 | YES |
| | | 2 | 1 4 8 15 19 20 22 37 39 44 ***9 18 26 29 49*** | 865 812 831 868 863 877 833 855 836 824 833 837 890 886 854 | 23 | YES |
| | | 3 | 6 7 8 9 11 16 20 21 35 47 ***12 25 30 39 42*** | 860 833 845 873 821 839 852 872 857 846 886 826 876 854 843 | 58 | YES |
| | | 4 | 3 11 13 27 35 39 42 46 47 50 ***10 20 21 32 50*** | 846 861 826 838 882 867 834 847 837 872 829 889 883 826 840 | 61 | YES |
| | | 5 | 5 7 10 16 22 29 30 33 42 44 ***6 13 20 37 47*** | 844 842 826 849 852 840 851 834 879 825 811 827 885 864 867 | 12 | YES |

[1] The sites indicated in ***bold italics*** are the 5 new sites that are specific to the daughter populations ($D_2$ and $D_3$), in each simulation run, in addition to the 10 sites of the ancestral population ($D_1$).

**Table 3.7**
**Mean ± s.d. of the Number of Sweeps to Convergence in 1000 Independent Simulation Runs for Synthetic Data set - 3**

| Population (Data Matrix) | Mean ± s.d. of the number of sweeps to convergence |
|---|---|
| Population 1($D_1$) | 45.76 ± 32.15 |
| Population 2($D_2$) | 33.25 ± 12.96 |
| Population 3($D_3$) | 32.86 ± 13.07 |

of E(S). This strategy may be computationally very expensive. Even if such a computationally expensive strategy is adopted, a measure to evaluate whether the drop in the value of E(S) for two consecutive values of p is "substantial" enough to stop the iterative algorithm. The search algorithm becomes computationally more feasible, with appropriate modifications, to provide the "best" value of p, when a range of values for p is specified, that is when $p_{min}$ and $p_{max}$ are specified with $p \in [p_{min}, p_{max}]$.

For any given value of the motif length $p \in [p_{min}, p_{max}]$, we can use the algorithm described for identifying motifs with a given motif length, and obtain the (maximum) value of E(S) given p, which we shall denote as E(S|p). We, therefore, calculate, $E(S|p_{min})$, $E(S|p_{min}+1)$, ... , $E(S|p_{max})$. Let $d(p_i)$ denote the value of $E(S|p_i)-E(S|p_i+1)$, $p_i \in [p_{min}, p_{max}-1]$. We now use the following statistical criterion to assess the significance of decrease in E(S|p) as the motif length (p) is increased, and stop with the smallest value of the motif length which satisfies this criterion.

$$\text{Let } \overline{d(p_i)} = \sum_{p_j=p_{min}}^{p_i-1} d(p_j)/(p_i - p_{min}),$$

and $$\sigma^2(p_i) = \sum_{p_j=p_{min}}^{p_i-1} \frac{[d(p_j)]^2}{p_i - p_{min}} - [\overline{d(p_i)}]^2 ; p_{min} < p_i \leq p_{max}.$$

If $|d(p_i)-\overline{d(p_i)}| > 2.\sigma^2(p_i)$, then we declare the decrease from $E(S|p_i)$ to $E(S|p_i+1)$ as significant, and stop with the motif length $p_i$. In the rare event of $E(S|p_{min}) = E(S|p_{min}+1) = \ldots = E(S|p_{i-1})$, we use the stopping criterion $E(S|p_{i-1}) > 2.E(S|p_i)$, and declare the length of the motif as $p_i$.

**Search for a Motif of Unknown Length using Synthetic Data**

We have studied the performance of the above algorithm for motif-finding when the length the motif is unknown using various synthetic data sets. The synthetic data sets were designed in a similar way as *Data Sets 1* and *2* described earlier. To study the behaviour of the algorithm, we sequentially increased the motif length from 1 ($= p_{min}$) to 15 ($= p_{max}$). In *Data Set 1*, except for the first 10 sites, elements in sites 11-50 were generated randomly. Hence, the exact motif frequency is unknown for motif-length > 10.

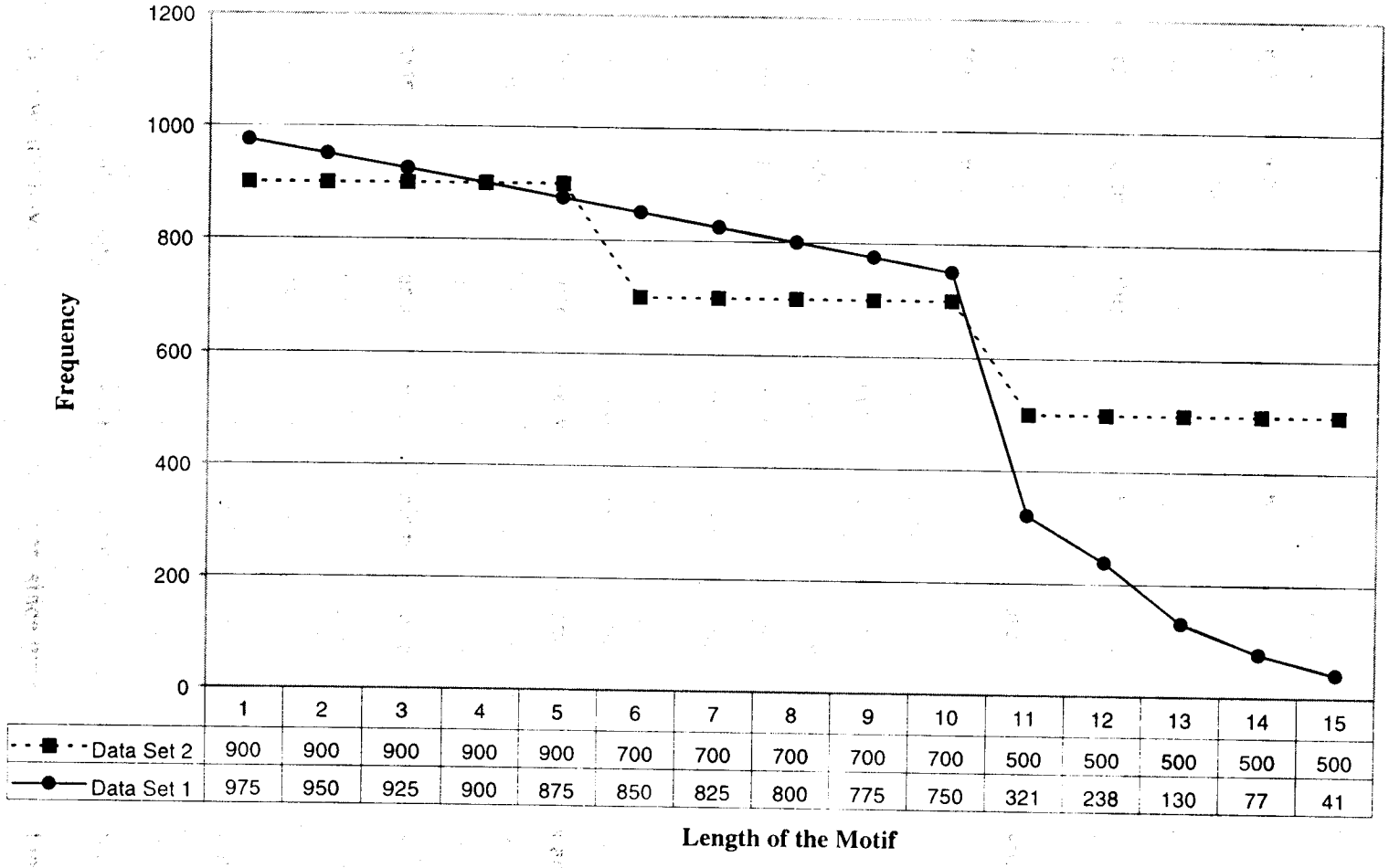
We have slightly modified the nature of of *Data Set 2* by changing the partition sizes, that is, the values of $\psi_1$, $\psi_2$ and $\psi_3$. We note that once the partition-sizes are chosen, the data set becomes deterministic. In Figure 3.1 we present the results of two different runs: one pertaining to *Data Set 1* another to *Data Set 2*. The motif length determined by our proposed algorithm in both cases is 10.

We first describe the results obtained using ***Data Set 2***. In this data set, we varied $\psi_2$ ($\geq 1$) and $\psi_3$ ($\geq 1$), such that $\psi_2 + \psi_3$ equaled K (K=5, 6,...,10), and $\psi_1 = 50-(\psi_2+\psi_3)$. In all cases, the length of the motif determined by our proposed algorithm was K with a frequency of 700. Figure 3.1 shows a specific combination with K=10 and $\psi_2 = \psi_3 = 5$. The motif length that the algorithm determined was 10 and hence the motif frequency was 700.

With respect to *Data Set 1*, the drop in motif frequency for each unit increase in length is the same (= 25) for the first 10 sites, and then there is a sudden drop in frequency when the motif length is 11. For this data-set, the algorithm determined the motif length (=10) in roughly 50% of the cases. In the remaining 50% of the cases, motif of the "correct" length was not identified. The reason for this is that we have proposed a conservative stopping criterion. That is, if $\sigma^2(p_i)=0$, then $E(Sl\, p_{i-1})$ must be greater than $2\times E(Sl\, p_i)$ to stop with a motif length of $p_i$. The choice of the constant 2 here is conservative and arbitrary. Since for the sites 11-50, 1s and 0s were randomly assigned (each with probability 0.5), there is no guarantee that $E(Sl\, p_{10})$ will be greater than $2\times E(Sl\, p_{11})$.

## Applications of the Algorithms to Real Data Sets

*Mitochondrial DNA Haplogroups M and U*: Extensive sequence data on the hypervariable segment 1 (HVS1) of the human mitochondrial (mt) DNA sampled from various global populations are now available in the public domain (www.hvrbase.org), including those generated by us and described in Chapter 2. The HVS1 region spans a set of 360 nucleotides, starting from nucleotide position (np) 16024 and ending at np 16383. On the basis of variations at specific nucleotide positions on the mtDNA outside




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Data Set 2 | 900 | 900 | 900 | 900 | 900 | 700 | 700 | 700 | 700 | 700 | 500 | 500 | 500 | 500 | 500 |
| Data Set 1 | 975 | 950 | 925 | 900 | 875 | 850 | 825 | 800 | 775 | 750 | 321 | 238 | 130 | 77 | 41 |

**Figure 3.1**
**Representative results pertaining to *Data Sets 1* and *2***

of the HVS1 region, mtDNA molecules have been classified into various haplogroups (HGs), such as M, U, etc. (Detailed definitions of HGs have already been provided in Chapter 2.) Because the HVS1 is a fast-evolving segment of the mtDNA, and because evolutionary antiquities and geographical substructuring of various HGs are different (as reviewed in Chapter 2), it is natural to expect "signature motifs" in the HVS1 region for samples belonging to the disjoint mtHGs. Some variants at specific HVS1 positions that are associated with some mtDNA HGs have already been identified (Macaulay et al. 1999; Quintana-Murci et al. 1999). The search for these variants is carried out, by visual inspection or by network analysis (Bandelt et al. 1995), after aligning sampled sequences in relation to the Cambridge Reference Sequence (CRS).

We have used the probabilistic search algorithms proposed here to identify motifs in some real data sets. As stated earlier, motif-search is carried out with respect to a reference sequence, and the motif length is unknown. In such a case, motif-search pertains to not only polymorphic sites in the set of sampled sequences, but a greater weightage needs to be given to those sequences that differ from the reference sequence at a larger number of nucleotide positions. This is because, one is generally interested in identifying motifs that are different from those present among individuals who carry the reference sequence, analogous to searching from "mutant" motifs that are different from "wild-type" motifs. Unless a greater weightage is given, motifs in which nucleotides included at the various positions of the motif may turn out to be the same as those present in the reference sequence. Thus, we are trying to choose such strings and nucleotides (at the sites included in the strings) at which the number of nucleotide differences compared to the reference sequence is "large", and then choose the string and nucleotides with the maximum frequency among them. For this purpose, given a specific string, $S_p$, of length p, we enumerate from the data all possible sequences of nucleotides, and calculate their frequencies, at the sites included in $S_p$. Among such sequences $\xi_l \mid S_p$ ($l$=1,2,...), we calculate their frequencies, $f_l$. We then calculate the number of mismatches of each of these sequences $\xi_l \mid S_p$ ($l$=1,2,...) with the reference sequences. Let $m_l$ denote the number of mismatches corresponding to $\xi_l \mid S_p$. If no weightage is given to $m_l$, then $E(S_p)$ would have been $\max_l(f_l)$. However, since we are interested in identifying

motifs that are different from the reference sequence, we need to take the $m_l$ values into account. To find a motif under such conditions, a greater weightage needs to be given to those sampled sequences that differ from the reference sequence. We have, therefore, used the following objective function

$$E(S_p) = \max_l ( f_l^{m_l} ).$$

We emphasize that this is not a unique choice for $E(S_p)$. It is possible to construct other objective functions under this situation, with the condition that such an objective function should be monotonically increasing in $f_l$ and $m_l$.

We have specifically focussed on mtDNA HGs M and U, because these are particularly relevant to India. The data that have been analyzed are those generated by us and described in Chapter 2. The total numbers of HVS1 sequences belonging to HGs M and U were, respectively, 338 and 115. For HG-M, the algorithm converged to p=4 and the corresponding string was S = (16223, 16270, 16319, 16352). The maximum frequency of this string, E(S), was 21 (= 6.21% of the total number of samples), and the nucleotides at the relevant nucleotide positions were T, T, A and C. The next most frequent string was (16223[T], 16274[A], 16319[A] and 16320[C]) with a frequency of 17(5.03%). For HG-U also the algorithm converged to p=4, and the motif was (16051[G], 16206[C], 16230[G], 16311[C]), with a frequency of 18 (= 15.65% of the total number of samples). The interpretations and implications of these motifs have already been discussed in Chapter 2, and are not repeated here.

## Summary

We have provided an efficient method of finding a motif, which we have defined as a set of nucleotides that occurs at a high frequency at multiple polymorphic DNA sites, not necessarily contiguous, in a set of DNA sequences each derived from an individual. The search for such motifs naturally arises in molecular evolutionary genetics and in genetic epidemiology, especially in case-control studies of diseases using the concept of disequilibrium mapping. We have provided two different probabilistic search algorithms when is motif length is known and is unknown. We have demonstrated and have extensively tested their efficiencies using both synthetic and real data sets. Our results indicate that the algorithms perform efficiently and are converge to the "correct" values in a small number of sweeps.

## Chapter 3: References


Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457-65

Balasubramanian S, Harrison P, Hegyi H, Bertone P, Luscombe N, Echoles N, McGarvey P, Zhang Z-L, Gerstein M (2002) SNPs on chromosomes 21 and 22 – analysis in terms of protein features and pseudogenes. Pharmacogenomics 3: 1-10

Bamshad MJ, Kivisild T, Watkins WS, Dixon MP, Ricker LE, Rao BB, Naidu M, Prasad BVR, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of Indian caste populations. Genome Res 11: 994-1004

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations. Genetics 141: 743-753

Bhandari D, Murty CA, Pal SK (1996) Genetic algorithm with elitist model and its convergence. Int J Pattern Recognit. Artif Intell 10: 731-747

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229-232

Gupta M, Liu JS (2003) Discovery of conserved sequence patterns using a stochastic dictionary model. JASA 98: 55-66

Keiler KC, Shapiro L (2001) Conserved promoter motif is required for cell cycle timing of *dnaX* transcription in *Caulobacter*. J Bacteriol 183: 4860-4865

Khani-Hanjani A, Lacaille D, Horne C, Chalmers A, Hoar DI, Balshaw R, Keown PA (2002) Expression of QK/QR/RRRAA or DERAA motifs at the third hypervariable region of HLA-DRB1 and disease severity in rheumatoid arthritis. J Rheumatol. 29:1358-65

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida,V., Scozzari, R., Bonné–Tamir, B., Sykes, B. and Torroni, A. (1999) The emerging tree of west Eurasians mtDNAs: a synthesis of control-region sequences and RFLPs. American Journal of Human Genetics 64, 232–249

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A and Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21: 1087-1091

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. Nat Genet 23: 437-441

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837

Schwaiger FW, Epplen JT (1995) Exonic MHC-DRB polymorphisms and intronic simple repeat sequences: Janus' faces of DNA sequence evolution. Immunol Rev. 143:199-224

Tateno Y, Ikeo K, Imanishi T, Watanabe H, Endo T, Yamaguchi Y, Suzuki Y, Takahashi K, Tsunoyama K, Kawai M, Kawanishi Y, Naitou K, Gojobori T (1997) Evolutionary motif and its biological and structural significance. J Mol Evol. 44 Suppl 1:S38-43.

# CHAPTER 4

## Estimating TMRCA from a Sample of DNA Sequences: A Comparison of Two Popular Statistical Methods

This Chapter has been accepted for publication in *Journal of Genetics* under the title "A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences," authored by Analabha Basu and Partha P. Majumder.

## Introduction

The assumption that underlies the statistical reconstruction of the evolutionary history of a set of contemporary populations is that new populations evolve over time by binary fission from ancestral populations. Looking backwards in time, therefore, a set of contemporary populations will coalesce pairwise at different points of time, until finally there is a coalescent event to the most recent common ancestor (MRCA) of all the populations. Such reconstruction can be done by using DNA sequence data generated from samples of individuals drawn from each of the contemporary populations under consideration. The two major features and parameters to be estimated from such data are (a) the topology of the coalescence events, and (b) the times of coalescence to common ancestors of the populations, including the time to MRCA (TMRCA). Both these and parameters are known to be affected by demographic scenarios that prevailed during the process of evolution (Hudson 1990; Nordborg 2001). The coalescent theory (Kingman 1982a, 1982b) provides a probabilistic framework and a method for reconstruction of evolution from DNA sequence data. The framework is simpler when one is dealing with a haploid, non-recombining, DNA molecule, such as mitochondrial DNA. Even with haploid DNA sequence data, estimating TMRCA based on a sample remains a major challenge. Saunders et al. (1984) have shown that although the TMRCA estimated from a sample can be different from the true TMRCA, the probability that the estimate will coincide with the true value is:

$$\frac{(n-1)}{(n+1)}\frac{(N+1)}{(N-1)} \cong \frac{(n-1)}{(n+1)},$$

where n is the sample size, N (>> n) is the population size (assumed to have been large and constant over evolutionary time. Thus, provided that we are dealing with numerically-large and temporally constant-size populations, even with a sample of 38 haploid DNA sequences (n), the probability of correctly estimating the true TMRCA is 0.95. Thus, the TMRCA of a sample is a reasonably good estimate of the TMRCA of the population (Saunders et al. 1984). Statistical methods have been developed to estimate TMRCA from a sample. However, the temporal constancy of population size is a crucial assumption underlying these methods. In practice, a population is expected to encounter

demographic pressures (such as, bottlenecks and expansions) resulting in violation of this assumption. The purpose of this study is to evaluate the impact of evolutionarily variable demographic scenarios on the estimates of TMRCA obtained by using two popular statistical methods (Templeton 1993; Bandelt et al. 1995; Sillard et al. 2000).

## Methodology

### *The Coalescent*

For completeness, we provide some key results of coalescent theory and briefly describe the two popular statistical methods.

The Kingman coalescent (Kingman 1982a, 1982b) is a probability model for the genealogical tree of a random sample of n genes drawn from a large population. A genealogical tree for a sample of size n=5 is depicted in Figure 4.1.

Time is measured continuously in the coalescent. The time $t_j$ during which the sample has *j* distinct lineages, $(2 \leq j \leq n)$ follows an exponential distribution with parameter j(j-1)/2 (Tajima(1983), Hudson(1991), Nordborg(2001)). The random variables denoting the times for different *j*s are independent. This description provides a close approximation to a range of population genetics models in which time is expressed in generations. An even larger class of models is approximated if a unit of coalescence time is interpreted as $N/\sigma^2$ generations, where $\sigma^2$ is the variance in an individual's number of offspring (Kingman 1982a). We shall assume $\sigma = 1$. We are primarily interested in the height of the tree $T_n$, i.e. the TMRCA.

It may be noted that,

$$T_n = t_n + t_{n-1} + \ldots + t_2 = \sum t_i, \text{ and}$$

$$E(t_j) = 2/\{j(j-1)\}$$

$$E(T_n) = 2(1-1/n)$$

$$Var(T_n) = 8\sum 1/j^2 - 4(1-1/n)^2 \rightarrow 4\pi^2/3 - 12 \approx 1.16.$$

In all the above expressions time is measured in units of N generations.

It is clear that as n increases, $E(T_n)$ very rapidly converges to 2. Also, $T_n$, the TMRCA has a large variance relative to the mean and this ratio does not reduce much by increasing the sample size.

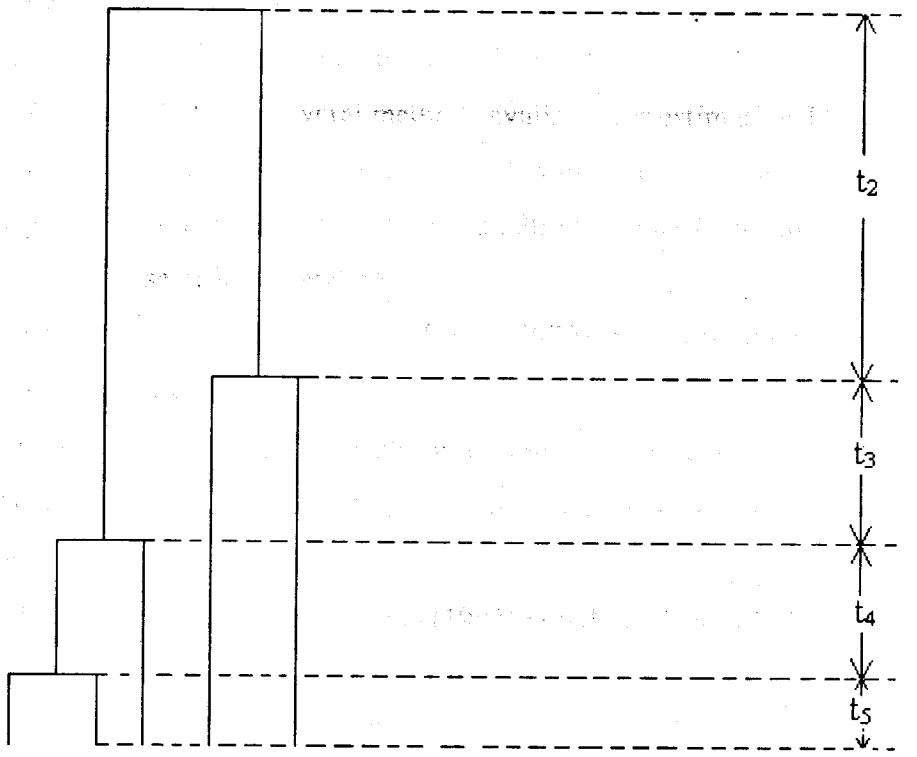



**Figure 4.1** Genealogical tree of a sample of 5 genes

The times at which mutations occur are modeled in the coalescent by assuming that these times form a Poisson process of constant rate $\mu$, where $\mu$ is the mutation rate per sequence per generation. This means that the number of mutations that may have accumulated on a branch of time-length $l$ is a realization from the Poisson distribution with mean $\mu l$. For DNA sequence data, if we assume that mutation rate has remained constant across sites and over time, then $\mu$ is equal to the sequence length times the mutation rate per site per generation.

Although there are several methods available for estimating TMRCA from a sample of DNA sequences (Fu and Li 1996; Tavare et al 1997) two methods are widely used (e.g., Mountain et al. 1995; Sillard et al. 2000) primarily because of conceptual simplicity and ease of interpretation.

Under the infinite sites model (Ewens 1979), all information in two DNA sequences is captured by the total number of segregating sites ($S_2$). Since, $E(S_2|T_2) = \theta T_2$ one approach of estimating $T_2$, which for a sample of two sequences is the TMRCA, by $S_2/\theta$. This and similar approaches (Hammer 1995) are not capable of utilising prior historical demographic information.

Using Bayes' Theorem, Tajima (1983) noted that if $S_2 = k$ then the distribution of $T_2$ is Gamma with parameters $1+k$ and $1+\theta$. In particular,

$$E(T_2|S_2=k) = (1+k)/(1+\theta)$$

$$Var(T_2|S_2=k) = (1+k)/(1+\theta)^2$$

Templeton(1993) considered the problem of estimating the TMRCA of n (>2) sequences by extending the analytical results that hold for n=2 and calculated the number of differences between each pair of sequences whose common ancestor is the root of the tree and then averaged these pairwise differences. He also observed that this value, $\hat{k}$, of k varied little over plausible reconstructed trees. He then substituted k by $\hat{k}$ in the previous equations for $E(T_2|S_2=k)$ and $Var(T_2|S_2=k)$. In a different study Hammer (1995) estimated the TMRCA for multiple sequences by substituting the largest value of k among all pairs in the previous equations. This is not a proper approach, because Donnelly and Kurtz (1997) have shown that the maximum number of differences

between a pair of sequences chosen from this set of n sampled sequences goes to infinity as n goes to infinity. This is true even when $T_n$ is bounded.

A popular alternative to the above procedures of estimating TMRCA, is to use median-joining network analysis (Bandelt et al 1995). In this analysis, a genealogy of n individuals is considered as an ultrametric tree, in which the lengths of links are scaled to time and each interior node corresponds to a coalescent event. If there are k ($\leq 2n-2$) links of lengths $t_1, t_2, \ldots, t_k$ time units and if the clade defined by the i-th link carries $n_i$ individuals (i= 1,2,...,k) then the coalescent time t can be expressed as

$$t = (n_1t_1 + n_2t_2 + \ldots + n_kt_k)/n.$$

If $\mu$ denotes the mutation rate, expressed as the expected number of (scored) mutations in a sequence segment per time unit, one may associate to the $i^{th}$ link a Poisson distributed random variable $X_i$ with parameter $\mu_i=t_i\mu$. The random variable $X=(n_1X_1+n_2X_2+\ldots+n_kX_k)/n$, has the expected value

$$E(X)=\{(n_1t_1 + n_2t_2 + \ldots + n_kt_k)/n\}\mu=t\mu$$

and variance

$$V(X)=\{(n_1^2t_1+n_2^2t_2+\ldots+n_k^2t_k)/n^2\}\mu,$$

assuming independence of $X_1, X_2, \ldots, X_k$.

## *Simulation method*

We evaluated the performance of these two methods for estimating the coalescent times from DNA sequence data. The data set consisted of nucleotide sequences from homologous segments of DNA sampled from different individuals. The data generated are similar to haploid nucleotide sequences, such as of the mtDNA hypervariable sequence 1 and 2 (http://www.hvrbase.org).

We have used a forward propagating algorithm to generate simulated DNA sequence data. In this algorithm a nucleotide sequence of a specified length and base composition is created by a multinomial random number generator with cell probabilities equal to the probabilities of the four bases. A completely homogeneous founding population of a given size is then formed by making the appropriate number of copies of the randomly generated nucleotide sequence. The founding population then evolves in accordance with the Wright-Fisher model (Ewens 1979), i.e. a new generation is formed

by sampling from the previous generation with replacement. The numerical size of the succeeding generations is controlled after the founding population is created. In this study we have considered two demographic scenarios: (a) constancy of population size over generations, and (b) exponential growth in size, allowing for variability in the growth parameter over generations. That is, when the size of a new generation is determined, we randomly selected the appropriate number of sequences from the gene pool of the previous generation with replacement. Then, using the assumed value of the mutation rate, we calculated the expected number of mutations per generation, and determined the number of new mutations to be introduced in each generation. If the expected number of mutations per generation is denoted as y, then we randomly chose and mutated [y] or [y]+1 sites, where [y] denotes the largest integer $\leq$ y. Choice between [y] or [y]+1 was made randomly by generating a random number $u$ from the uniform [0,1] distribution, where [y] was chosen if $u$ was less than y–[y]. Suppose there are $N_t$ individuals in generation $t$, each with data on a sequence of L nucleotide sites. To introduce a new mutations in generation t, a site was chosen with probability $1/(N_t \times L)$ and mutated. If $x_1$ is one such observation, then the mutation is introduced at the nucleotide position $((x_1/L)-[x_1/L])\times L$ of the $[x_1/L]$-th individual. While introducing the mutation, we did not consider any prior information on mutational histories of the site or the individual, thus allowing for parallel, recurrent and back mutations to occur. This process is thus repeated for a stipulated number of generations. The population thus generated was treated as the present population and a random sample of size n was drawn without replacement. This sample of n sequences then was used to estimate the TMRCA of the population. The estimated TMRCA was compared to the actual number of generations used in the simulation.

Simulation Parameters: Since estimates of TMRCA can be affected by various parameters, we have investigated the effects of variation in four crucial parameters. These are:

(1) The number of bases (L) of the nucleotide sequence; we have used two different values of L – 200 and 400.

(2) Variability in population size over generation, which was introduced through a parameter $\alpha$. We have used an exponential growth model. In this model, if $N_t$ denotes the

population size in generation t, then $N_{t+1} = N_t e^{\alpha}$. In order that $N_{t+1}$ is an integer, we have chosen either $[N_t e^{\alpha}]$ or $([N_t e^{\alpha}]+1)$. Choice between $[N_t e^{\alpha}]$ or $[N_t e^{\alpha}]+1$ was made randomly by generating a random number $u$ from the uniform [0,1] distribution; $N_{t+1} = [N_t e^{\alpha}]$ was chosen if $u$ was $< (N_t e^{\alpha}) - [N_t e^{\alpha}]$; otherwise $N_{t+1} = [N_t e^{\alpha}]+1$ was chosen. We have used three different values of $\alpha$ – 0, 0.001, 0.005.

(3) The number of generations (g); three different values of g – 250, 500 and 1000 – were used.

(4) Mutation rate ($\mu$); two values were used which were $\mu = 10^{-5}$/site/generation and $\mu = 5\times10^{-5}$/site/generation. These values roughly correspond to the observed rates in human autosomal and mitochondrial hypervariable segment-1, respectively. We note that although the relevant theoretical equations are functions of $N\mu$, we have varied N and $\mu$ independently to study the effect of parallel and back mutations, which are possibly introduced when $\mu$ is large.

## Results and Discussion

Simulated data were generated using different combinations of the parameter values stated above. For each simulated data set, estimation of TMRCA was carried out using two different methods (Templeton 1993, Bandelt et al. 1995). TMRCA was estimated from a sample of n=100 sequences. Since both estimation procedures crucially depend on the number of segregating sites, for a data set to be 'informative', the sample of sequences must have at least two segregating sites. We encountered non-informative data sets in our simulation runs, especially when g and $\mu$ were both small. Our comparisons are all based on 100 informative data sets; that is, 100 data sets each of 100 sequences containing at least two segregating sites. Two typical simulated data sets and the corresponding median-joining networks are given in Figures 4.2 and 4.3.

We first note that a large number of simulation runs was required to generate 100 informative data sets, because often the generated data set did not contain even two segregating sites. This number was particularly large when either g or $\mu$ was small. For the MJN analysis, a further problem was encountered for an informative data set that had a single segregating site. For such a data set, while it was possible to calculate $\hat{k}$, it was

| Total number of segregating sites = 3 | | |
|---|---|---|
| Number of distinct sequences = 4 | | |
| Sequence No. | Sequence | Frequency |
| 1 | CGC | 94 |
| 2 | CAC | 3 |
| 3 | TGC | 2 |
| 4 | CGA | 1 |

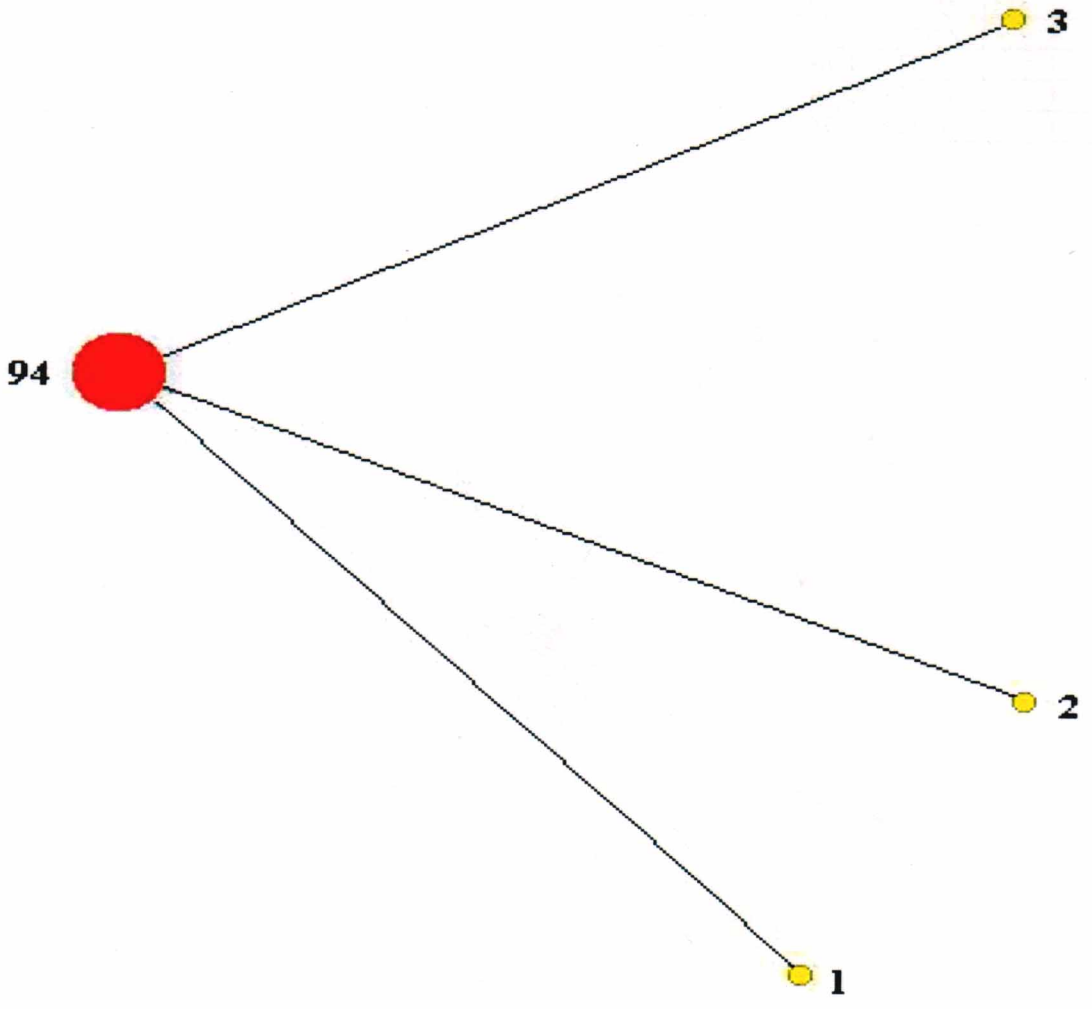



**Figure 4.2**
A typical simulated data set with parameters L=400, $\alpha$=0.0, g=500, $\mu=5\times10^{5}$ and its corresponding median joining network. The red circle denotes the ancestral sequence while the numbers beside the circles are the frequencies of individuals with that sequence.

| Total number of segregating sites = 9 | | |
|---|---|---|
| Number of distinct sequences = 10 | | |
| Sequence No. | Sequence | Frequency |
| 1 | ACCGTTGAC | 80 |
| 2 | ACCCTTGAC | 4 |
| 3 | ACCGTTGAT | 5 |
| 4 | ACCGTTTAC | 4 |
| 5 | AGCGTTGAC | 1 |
| 6 | GCCGTTGAC | 1 |
| 7 | ACCGTGGAC | 2 |
| 8 | ACTGTTGAC | 1 |
| 9 | ACCGTTGGC | 1 |
| 10 | ACCGATGAC | 1 |

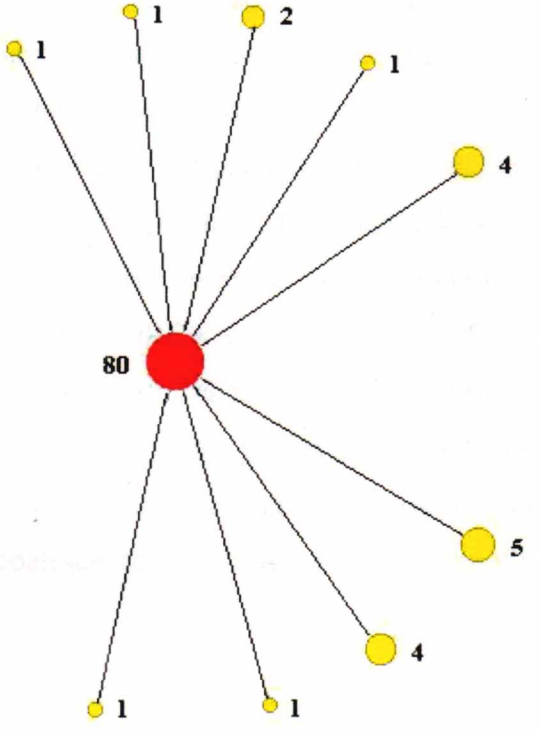



**Figure 4.3**
A typical data set with parameters L=400, $\alpha$=0.005, g=500, $\mu=5\times10^{5}$ and its corresponding median joining network. The red circle denotes the ancestral sequence while the numbers beside the circles are the frequencies of individuals with the specific sequence

not possible to draw the network (using the MJN software) and, therefore, to estimate TMRCA from the MJN. We had to discard such data sets from the MJN analysis. To keep the results comparable, we, however, generated 100 informative data sets on which both methods of estimating TMRCA could be implemented.

Our results are summarized in Table 4.1. It is evident from Table 4.1 that the standard deviations (SDs) of the TMRCA estimates were very large, irrespective of the parameter values used in the simulation. Generally, both methods underestimated the true TMRCA, except for short sequence lengths (L=200, 500) and a short evolutionary time (g=250, 500) with a low mutation rate ($\mu = 10^{-5}$). However, the means of the estimated TMRCA values were not significantly different from the true values because of the large SDs. The correlation coefficient of the TMRCA estimates by the two methods is large for all sets of simulation parameter values. Thus, both methods seem to be rather unreliable in practice and it is difficult to choose between the two.

The means of the estimated TMRCA values for most combinations of simulation parameter values decreased as the mutation rates increased. This is because with a higher mutation rate there is a higher probability of parallel and back mutations, especially when the lengths of the sampled sequences were short. Both methods were rather insensitive to the population growth parameter ($\alpha$), and there was no consistent trend with respect to $\alpha$ of the either the mean values of the TMRCA estimates or the SDs, although the SDs in many cases decreased with increase in $\alpha$. The frequency distributions of the TMRCA estimates (Figures 4.4 and 4.5) were all highly positively skewed with a very long upper tail for both methods. Our results indicate that in practice considerable caution needs to be exercised in interpreting coalescence times estimated by either of these two methods, which are quite popular.

**Table 4.1**

Mean ± s.d. values of TMRCA estimated by Templeton's ($T_1$) and Median-joining-network ($T_2$) methods for various sets of parameter values

| Sequence Length (L) | No. of generation (g) | Growth Rate ($\alpha$) | $\mu = 10^{-5}$ | | | $\mu = 5 \times 10^{-5}$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | $\hat{T}_1 \pm$ s.d. | $\hat{T}_2 \pm$ s.d. | Correlation | $\hat{T}_1 \pm$ s.d. | $\hat{T}_2 \pm$ s.d. | Correlation |
| 200 | 250 | 0 | 361±316.1 | 398±313.9 | 0.95 | 216±132.4 | 178±123.7 | 0.98 |
| | | 0.001 | 363±338.5 | 355±329.5 | 0.89 | 198±135.8 | 108±114.4 | 0.88 |
| | | 0.005 | 346±278.2 | 276±234.1 | 0.86 | 187±106.1 | 95±75.6 | 0.88 |
| | 500 | 0 | 531±503.1 | 583±620.0 | 0.93 | 374±179.8 | 387±262.2 | 0.80 |
| | | 0.001 | 507±412.7 | 446±414.2 | 0.87 | 326±178.4 | 298±191.7 | 0.94 |
| | | 0.005 | 428±346.2 | 251±306.4 | 0.85 | 394±139.6 | 303±112.0 | 1.00 |
| | 1000 | 0 | 908±808.0 | 1167±1080.0 | 0.93 | 629±318.6 | 639±376.2 | 0.94 |
| | | 0.001 | 825±694.5 | 890±805.4 | 0.96 | 498±322.8 | 505±376.7 | 0.94 |
| | | 0.005 | 858±394.2 | 644±312.5 | 1.00 | 910±178.5 | 696±153.0 | 0.99 |
| 400 | 250 | 0 | 218±212.3 | 374±422.5 | 0.93 | 205±91.0 | 342±167.5 | 0.96 |
| | | 0.001 | 231±177.4 | 305±294.0 | 0.82 | 199±89.0 | 217±139.0 | 0.88 |
| | | 0.005 | 225±167.5 | 288±345.2 | 0.81 | 206±74.7 | 215±103.5 | 0.89 |
| | 500 | 0 | 429±370.7 | 833±905.4 | 0.92 | 387±162.3 | 701±331.7 | 0.98 |
| | | 0.001 | 439±306.8 | 962±774.3 | 0.89 | 351±150.0 | 620±300.6 | 0.98 |
| | | 0.005 | 375±201.4 | 590±317.6 | 1.00 | 437±124.8 | 681±215.8 | 0.99 |
| | 1000 | 0 | 637±464.3 | 1331±1031.4 | 0.95 | 611±219.0 | 1190±518.0 | 0.94 |
| | | 0.001 | 751±527.3 | 1520±1190.5 | 0.91 | 606±204.8 | 1129±444.3 | 0.96 |
| | | 0.005 | 834±287.8 | 1262±469.2 | 1.00 | 912±81.6 | 1375±145.6 | 0.99 |

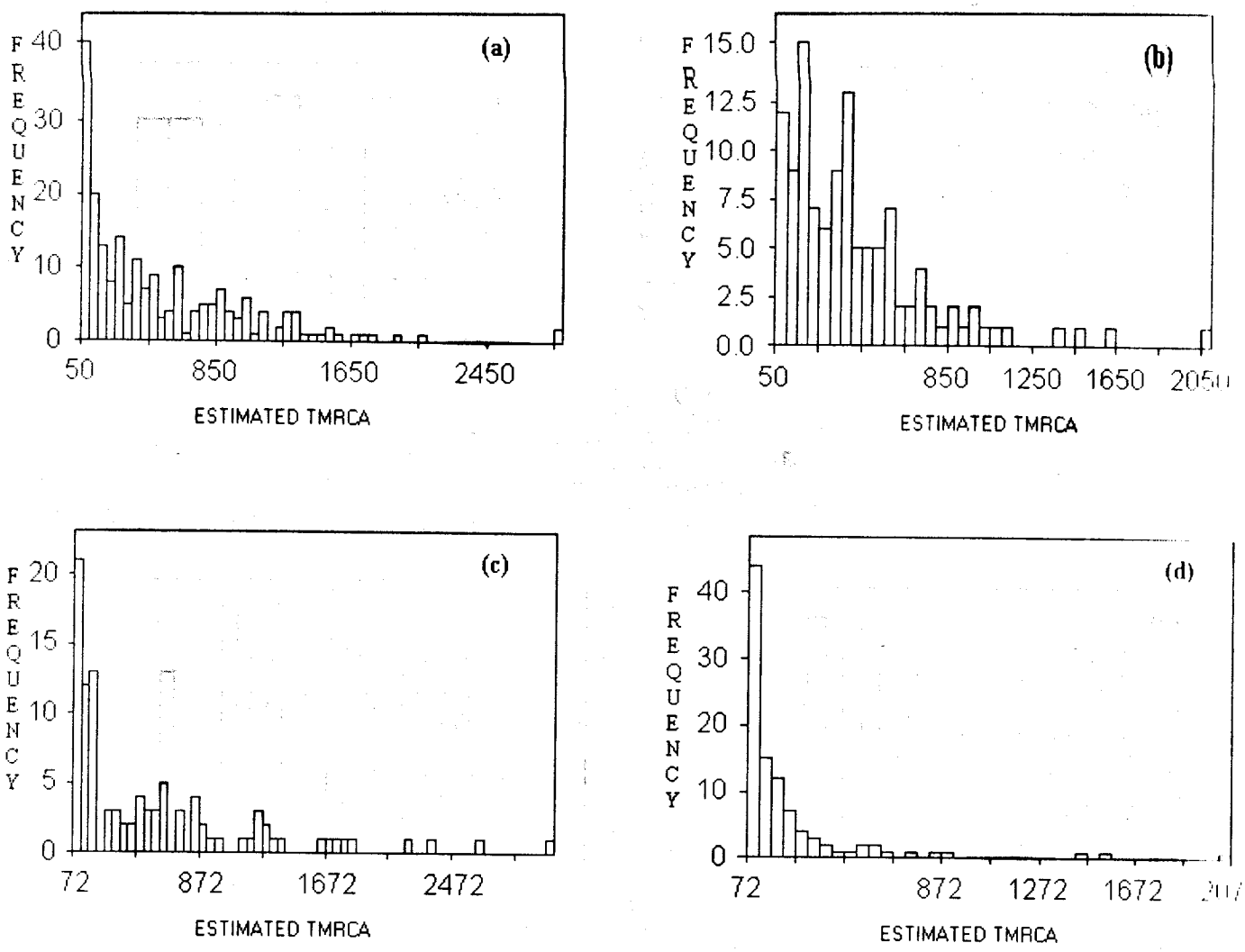



**Figure 4.4** Frequency distributions of TMRCA estimated by two methods on simulated DNA sequence data, with simulation parameters L=200 nucleotides, $\mu = 10^{-5}$/site/ generation and g=500 generations (marked with an arrow on the X-axis), comprising of 100 replications of each data set. (a) $\alpha$=0, Templeton's estimation method, (b) $\alpha$=0.005, Templeton's estimation method, (c) $\alpha$=0, MJN estimation method, (d) $\alpha$=0.005, MJN estimation method.

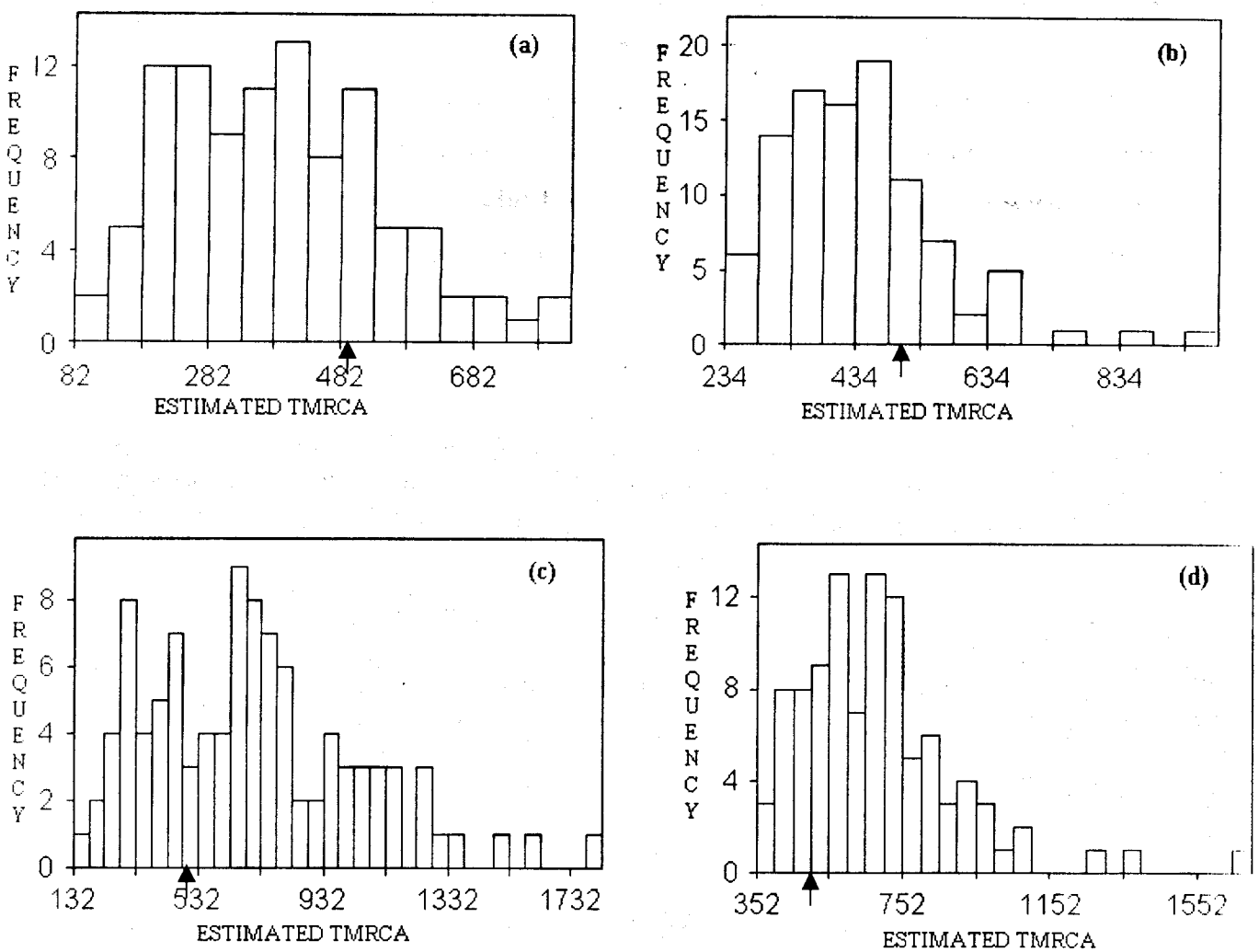



**Figure 4.5** Frequency distributions of TMRCA estimated by two methods on simulated DNA sequence data, with simulation parameters L=400 nucleotides, $\mu = 5 \times 10^{-5}$/site/ generation and g=500 generations (marked with an arrow on the X-axis), comprising of 100 replications of each data set. (a) $\alpha$=0, Templeton's estimation method, (b) $\alpha$=0.005, Templeton's estimation method, (c) $\alpha$=0, MJN estimation method, (d) $\alpha$=0.005, MJN estimation method.

## Summary

We have compared two statistical methods of estimating the TMRCA from a sample of DNA sequences, which have been proposed by Templeton (1993) and Bandelt et al. (1995). Monte-Carlo simulations were used for generating DNA sequence data. Different evolutionary scenarios were simulated and the estimation procedures were evaluated. We have found that for both methods (i) the estimates are insensitive to demographic parameters, and (ii) the standard deviations of the estimates are too high to be reliably used in practice.

## Chapter 4: References


Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations. Genetics 141: 743-753

Donnelly P, Kurtz TG (1996) The asymptotic behaviour of an urn model arising in population genetics. Stoch. Proc. Appl. 64: 1-16

Ewens WJ (1979) *Mathematical population genetics*. Springer, New York

Fu Y-X, Li WH (1997) Estimating the age of the common ancestor of a sample of DNA sequences. Mol. Biol. Evol. 14: 195-199

Hammer MF (1995) A recent common ancestry for human Y chromosomes. Nature 378: 376-378

Hudson RR (1990) Gene genealogies and the coalescent process. In Futuyma D, Antonovics J (eds) *Oxford Surveys of Evolutionary Biology*, 7: 1-44. Oxford: Oxford University Press

Kingman JFC (1982a) On the genealogy of large populations. J. Appl. Prob. 19A: 27-43.

Kingman JFC (1982b) The coalescent. Stoch. Proc. Appl. 13: 235-248

Mountain JL, Hebert JM, Bhattacharyya S, Underhill PA, Ottolenghi C, Gadgil M et al. (1995) Demographic history of India and mtdna sequence diversity. Am J Hum Genet 56: 979-992

Nordborg M (2001) Coalescent theory. In: Balding D, Bishop M, Cannings C (eds.) *Handbook of Statistical Genetics*. Pp. 341-385, Wiley, Chichester, UK

Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67: 718-726

Saunders IW, Tavare S, Watterson GA (1984) On the genealogy of nested subsamples from a haploid populaton. Adv Appl Prob 16: 471-491

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437 -460

Tavare S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505-518

Templeton AR (1993) The "eve" hypothesis: a genetic critique and reanalysis. Amer Anthrop 95: 51-72

# CHAPTER 5

## A Statistical Method to Estimate Relative Times of Divergence of Populations from a Common Ancestor

## Introduction

Reconstruction of evolutionary histories of populations is often done from data on DNA sequences from samples of individuals drawn from these populations using phylogenetic methods (Fitch and Margoliash 1967; Goodman et al. 1971; Holmquist 1972; Saitou and Nei 1987). The two problems in phylogenetic analysis are (a) estimation of topology, and (b) estimation of branch lengths. It is known from theoretical studies and extensive simulations (Tateno et al 1982; Nei, Tajima and Tateno 1983) that correct estimation of topology is easier than estimation of branch lengths with low error. In Chapter 4, we have provided statistical evidence that even the estimate of the time to the most recent common ancestor (TMRCA) can be poor, when popular estimation methods, such as those of Templeton (1993) and Bandelt et al. (1995), are used. Additionally, past demographic histories, such as whether the population size has remained constant or whether the population has passed through a bottleneck; affect the phylogenetic relationships among DNA sequences, particularly branch lengths (Hudson 1990; Nordborg 2001). Even to estimate TMRCA, prior knowledge, or minimally some assumptions, of a population parameter $\theta=4N\mu$ ($\mu$=mutation rate/site/generation) is required, which is often unknown (Chakraborty 1977).

The purpose of this Chapter is to propose a statistical method to efficiently estimate *relative* branch lengths, which is often sufficient for evolutionary inferences. This method does not require prior knowledge or make any assumptions on $\theta$.

## Statistical Methodology

Consider a sample of $n$ haploid DNA sequences. Let $k_{ij}$ denote the number of mismatches between the *i-th* and *j-th* sequences ($1 \le i < j \le n$); that is, $k_{ij}$ denotes the number of nucleotide positions at which the *i-th* and *j-th* sequences differ. Under the infinite sites model (Ewens 1979), these $k_{ij}$ differences arose after the two sequences diverged from a common ancestor. If $t_{ij}$ denotes the time since divergence of these two

sequences, $i$ and $j$, from their common ancestor, and if $\mu$ denotes the mutation rate per site per generation, then

$$E(k_{ij}) = 2\mu\, t_{ij}\,.$$

Now, to fix ideas, consider $n = 4$. Suppose, without loss of generality, the evolutionary relationships among the 4 sequences and their divergence times are as shown in Figure 5.1. As is seen from this figure,

$$E(k_{12}) = 2\mu\, t_4\,,\ E(k_{34}) = 2\mu\,(t_4 + t_3) \text{ and } E(k_{14}) = 2\mu\,(t_4 + t_3 + t_2)$$

Thus, we can write,

$$E\begin{bmatrix} k_{12} \\ k_{13} \\ k_{14} \\ k_{23} \\ k_{24} \\ k_{34} \end{bmatrix} = 2\mu \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} t_4 \\ t_3 \\ t_2 \end{bmatrix}$$

or,

$$E(\boldsymbol{k}) = 2\,\mu\,\boldsymbol{B}\,\boldsymbol{t}\,,$$

where $\boldsymbol{k}$, $\boldsymbol{B}$ and $\boldsymbol{t}$ have obvious definitions. This equation is easily generalizable to data on pair-wise mismatches among $n$ sequences.

If $t^* = 2\mu\, t$, then an ordinary least squares estimator of $t^*$ is

$$\hat{t}^* = (\mathrm{B'B})^{-1}\mathrm{B'}\boldsymbol{k}\,,$$

where $\boldsymbol{B}$'= transpose of $\boldsymbol{B}$. The problem with this estimator is that estimates of individual components of $\hat{t}^*$ may be $\leq 0$, when in fact times of divergence can not be negative. To ensure that estimates of individual components of $\hat{t}^*$ are $\geq 0$, optimization has to be done on a restricted domain $t^* \geq 0$. To solve a related problem, Chakraborty (1977) had also used the least-squares method, but his approach is significantly different from that proposed by us.

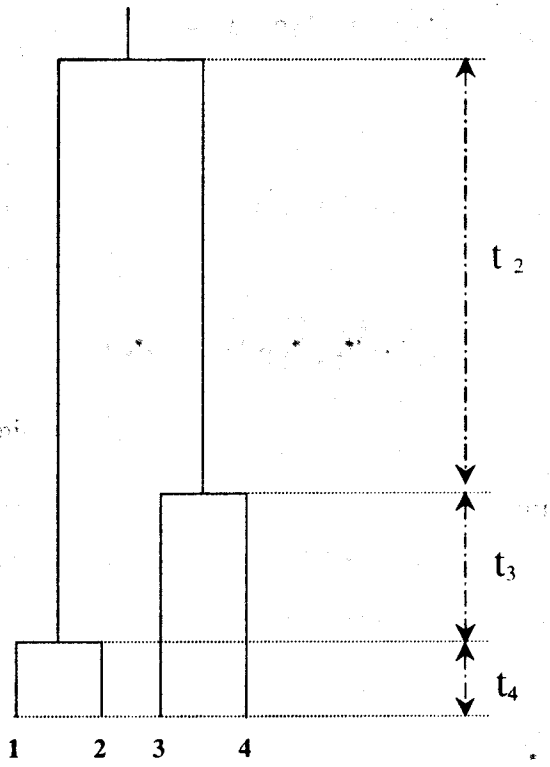



**Figure 5.1**
Genealogical tree of a sample of 4 DNA sequences (individuals)

We have used a quadratic programming (QP) approach to carry out optimization on the restricted domain $t^* \geq \mathbf{0}$. The error sums of squares under the model $k = \mathrm{B}t^* + \varepsilon$ is:

$$\begin{aligned} \varepsilon' \varepsilon &= (k - \mathrm{B}t^*)'(k - \mathrm{B}t^*) \\ &= k'k - 2k'Bt^* + t^{*'}(B'B)t^*. \end{aligned}$$

We need to minimize $\varepsilon' \varepsilon$ with respect to $t^* \geq \mathbf{0}$. Now, minimizing $\varepsilon' \varepsilon$ is equivalent to minimizing

$$f(t^*) = -2k'Bt^* + t^{*'}(B'B)t^*,$$

where $B'B$ positive semi-definite.

Therefore, $f(t^*)$ is convex (Hadley 1964), and hence any local minimum is a global minimum of $f(.)$. It is interesting to note that if the ordinary least squares estimate is in the feasible region (i.e., $t^* \geq \mathbf{0}$), then that estimate is the same as the one obtained by minimizing $f(.)$.

Our problem, therefore, is to minimize $f(t^*)$, subject to $t^* \geq \mathbf{0}$. The inequality constraints may be converted into equations by subtracting an appropriate non-negative 'surplus' variable. Thus, to satisfy the non-negativity condition, let $S_i^2$ $(\geq 0)$ $(i=2,3,\ldots,n)$ be the surplus quantity subtracted from the *i-th* constraint $t^*_i \geq 0$. Denote the vector of surplus variables as

$$S^2 = (S_2^2, S_3^2, \ldots, S_n^2)' \text{ and } S = (S_2, S_3, \ldots, S_n)'.$$

Thus, the function with a Lagrangian multiplier can be rewritten as

$$L(t^*, S, \lambda) = f(t^*) - \lambda [t^* - S^2]$$

Now,

$$\frac{\partial L}{\partial t^*} = \mathbf{0} \Rightarrow \lambda = \frac{\partial f}{\partial t^*}.$$

Obviously, $\lambda$ measures the rate of variation of $f$ with respect to $t^*$. Since the unconstrained minimum of $f$ is $\leq$ the constrained minimum, $\lambda$ should be non-negative.

Therefore, $\lambda = \frac{\partial f}{\partial t^*} \geq 0$. Taking partial derivatives of $L$ with respect to $t^*$, $S_i$ and $\lambda$, we obtain the following three equations,

$$\frac{\partial L}{\partial t^*} = \frac{\partial}{\partial t^*} f(t^*) - \lambda = 0 \quad \ldots \quad (1)$$

$$\frac{\partial L}{\partial S_i} = -2 \lambda_i S_i = 0 \quad (i=2,3,\ldots,n) \quad \ldots \quad (2)$$

$$\frac{\partial L}{\partial \lambda} = -[(t^*) - S^2] = 0 \quad \ldots \quad (3)$$

From equation (2), we deduce that if $\lambda_i$ is strictly greater than zero, then $S_i^2 = 0$. Also, if $S_i^2 > 0$, $\lambda_i = 0$. Now, if $\lambda_i$ is the i-th component of $\frac{\partial f}{\partial t^*}$, then it implies that $t_i^*$ does not affect the value of $f$. From equations (2) and (3), it follows that $\lambda_i t_i^* = 0$ ($i = 2, 3, \ldots, n$). Therefore, the Kuhn-Tucker conditions (Hadley 1964) necessary for $t^*$ and $\lambda$ to be a stationary point of the minimization problem are satisfied. The sufficient condition for global minimization, as has been mentioned earlier, is the convexity of $f(t^*)$, which is ensured because $\boldsymbol{B}'\boldsymbol{B}$ is positive definite. Thus, $f(t^*)$ satisfies both the Kuhn-Tucker necessary and sufficient conditions for existence of a global minimum. We used the MATLAB program MINQ (Neumaier 1998) to obtain the value of $t^*$, which minimizes $f(t^*)$.

Using quadratic programming (QP) we can get a numerical estimate of $t^* = 2\mu t$, and therefore $t_i^* = 2\mu t_i$ ($i=2,3,\ldots,n$). To obtain estimates of relative times of divergence, which are independent of $\mu$ (since the multiplicative constant $2\mu$ cancels out from both the numerator and the denominator) and hence do not require prior knowledge of $\mu$, we propose the natural estimator $\frac{\hat{t}_i^*}{\hat{t}_j^*}$ where $i=2,3,\ldots,j-1$ and $j=3,4,\ldots,n$. Under neutrality and constant population size, the expected value of $t_i$ is known. Consider a population that has evolved at constant growth rate (*e.g.*, an exponentially growing population with a growth

rate of $\alpha$ per generation). Looking backward in time, the effective population size $t$ generations ago ($N_t$) decreases as $t$ increases. As the probability of occurrence of a coalescent event in generation $t$ is $1/N_t$, the decrease in $N_t$ with respect to $t$ subsequently increases the probability of a coalescent event in each generation. This phenomenon, therefore, reduces the expected value of $t_j$, which becomes smaller in magnitude as we move backward in time. $E(t_2)$ is, therefore, affected to a greater degree than $E(t_n)$. Consequently, $E(t_2/t_n)$ is much smaller for an exponentially (with parameter $\alpha$) growing ($\alpha>0$) population than a population whose size has remained constant over time. As is evident, the scenario gets reversed when $\alpha$ is $< 0$, *i.e.*, when we consider a population which has exponentially decreased ($\alpha<0$) in size over time.

A similar pattern is also expected when we consider a population that has passed through a recent bottleneck. The behaviour of coalescence-time, considered as a random variable, in a population that has passed through a bottleneck is similar to a population with a constant effective size until the time of bottleneck. As we look backwards in time, due to the sudden increase in population size prior to the bottleneck, the expected time for the coalescent events occurring before the bottleneck increases. Thus $E(t_2 \mid \text{bottleneck}) >> E(t_2 \mid \text{constant population size})$. This results in a sudden increase in ratios, such as $E(t_2/t_n)$. These systematic trends, caused by various demographic histories of a population, will be reflected in the estimates of the different ratios of coalescent times.

## Assessment of Properties of the Estimator by Simulation

When the effective size of a population has remained constant over time, $t_i$s are independent and exponentially distributed (Hudson 1990; Nordborg 2001) with expectation $1/i(i\text{-}1)$ (after suitable adjustment of the time scale as 2N generations = 1 time unit for haploids, and 4N generations = 1 time unit for diploids). Under this assumption $v = \frac{t_i}{tj}$ has the following probability density function:

$$f(v) = \frac{\frac{i(i-1)}{j(j-1)}}{\left\{1 + \frac{i(i-1)}{j(j-1)} v\right\}^2}$$

The distribution of $v$ does not have a finite expectation in the range $(0, \infty)$. Since $t_i$ and $t_j$ are independent, using the fact that the arithmetic mean of a positive valued random variable is larger than the reciprocal of its harmonic mean, we can assert that

$$E\left(\frac{t_i}{t_j}\right) \geq \frac{E(t_i)}{E(t_j)} = \frac{j(j-1)}{i(i-1)}.$$

A comparison with the theoretical mean of $t_i/t_j$, to test the bias of the ratio estimator proposed by us is, therefore, not possible. Consequently, we have studied the properties of our estimator empirically. We have carried out extensive Monte-Carlo coalescent simulations to study the behavior of the estimator. Using the coalescent simulation program ***ms.c*** (Hudson 2002). We generated 1000 replicates of samples, each of size 5, under the constant effective population size assumption. The number of segregating sites for each set of 1000 samples was fixed. We have considered 3 different cases, when the number of segregating sites were 30, 40 and 50. From the coalescent program we collected the values of $t_n$ ($n$=2,3,4,5), *i.e.*, the time intervals in the genealogy when there were exactly $n$ lineages. From these values it was possible to estimate the empirical mean and distribution of $t_i/t_j$. Similarly, from the data sets generated by the program, we calculated $\hat{t}_i^*$ $s$ ($i$=2,3,4,5) using the proposed method. From the estimated $\hat{t}_i^*$ values, we calculated and plotted the distributions of the estimator $\hat{t}_i^* / \hat{t}_j^*$ ($i$ = 2,3,...,$j$-1; $j$ = 3,...,$n$). Using the 2-sample Kolmogorov-Smirnov test to compare the actual distribution of $t_i/t_j$ and the empirical distribution of the estimator, we have found that the hypotheses of the equality of the distributions were accepted at the 5% level of significance for all $i$ = 2,3,...,$j$-1; $j$=3,...,$n$ . To investigate the bias of the estimator, we carried out a non-parametric test of equality of the means of the distributions. The null hypothesis of the equality of the means was always accepted with a p-value > 0.6. Thus

the proposed estimator does not appear to be biased in any systematic way. Henceforth, we shall denote our estimator as $\hat{t}_i / \hat{t}_j$ instead of $\hat{t}_i^* / \hat{t}_j^*$.

We then generated samples, using the same coalescent program, under different demographic scenarios. We generated samples of sequences from 1000 independent populations each with a history of exponential growth. We have considered two different growth rates. We also generated samples from 1000 populations, each of which had undergone a recent bottleneck. We have considered two different cases by varying the time (in the past) of occurrences of the bottleneck and the magnitudes of the bottleneck. We also generated samples form a population that has experienced exponential expansion after the bottleneck. Under each scenario, we calculated the mean value of $\hat{t}_i / \hat{t}_j$ over the 1000 runs. Expectedly, the mean values of the exponentially growing populations were drastically smaller than those of the constant effective size populations, which in turn were smaller than those of the populations which underwent a bottleneck. These results are presented in Tables 5.1, 5.2 and 5.3.

We also expected that the distribution of $\hat{t}_i / \hat{t}_j$ for a population with a history of exponential growth to be stochastically smaller than that of a population which has remained in constant effective size, and that this distribution in turn, to be stochastically smaller than that of a population with a history of recent bottleneck. To test the above expectations we compared the distribution of $\hat{t}_i / \hat{t}_j$ obtained form relevant populations generated under different demographic history scenarios.

The distribution of $\hat{t}_i / \hat{t}_j$ for different possible choices of ***i*** and ***j*** under the constant effective population size model served as a null model to detect population size changes. As is popular in literature, this also serves as a null model for neutrality tests (Tajima 1989; Fu and Li 1993). As was expected, the distributions greatly varied (Tables 5.4, 5.5 and 5.6) for various numbers of segregating sites. For population that did not retain a constant effective size, the distributions of ratios of $t_j$s close to the MRCA showed greater deviation (as is seen from the p-values and also from the Kolmogorov-Smirnov (KS) statistic from the null, than for ratios of more recent divergence times, such

**Table 5.1**

Empirical Mean Values of $\hat{t}_i / \hat{t}_j$ under Different Demographic Scenarios when the Number of Segregating Sites = 50

| Statistic | Observed Empirical Mean under various Demographic History Scenarios | | | | | | $\frac{E(t_i)}{E(t_j)}$ |
|---|---|---|---|---|---|---|---|
| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 | Scenario 6 | |
| $\hat{t}_2/\hat{t}_5$ | 11.76 | 4.94 | 4.26 | 14.69 | 21.74 | 8.72 | 10.00 |
| $\hat{t}_3/\hat{t}_5$ | 5.09 | 3.51 | 3.28 | 7.03 | 7.30 | 3.25 | 3.33 |
| $\hat{t}_4/\hat{t}_5$ | 3.02 | 2.67 | 2.68 | 3.13 | 2.57 | 2.43 | 1.67 |
| $\hat{t}_2/\hat{t}_4$ | 8.65 | 4.80 | 3.46 | 12.27 | 18.06 | 10.36 | 6.00 |
| $\hat{t}_3/\hat{t}_4$ | 3.66 | 3.56 | 2.95 | 4.89 | 5.93 | 3.03 | 2.00 |
| $\hat{t}_2/\hat{t}_3$ | 7.87 | 3.84 | 3.26 | 16.36 | 13.11 | 11.15 | 3.00 |

Scenario 1: Constant effective population size

Scenario 2: Exponentially growing population: Population size was 1/20$^{th}$ of the present size 2N generations back

Scenario 3: Exponentially growing population: Population size was 1/100$^{th}$ of the present size N generations back

Scenario 4: Population with a history of bottleneck: Population size was 10 times the present size N generations back

Scenario 5: Population with a history of bottleneck: Population size was 5 times the present size N/2 generation back

Scenario 6: Population size was constant and 10 times the present size upto N/2 generations back. Then the size reduced to 1/10$^{th}$ of the present size in a single generation, after which it grew exponentially to the present size, N.

**Table 5.2**

Empirical Mean Values of $\hat{t}_i / \hat{t}_j$ under Different Demographic Scenarios when the Number of Segregating Sites = 40

| Statistic | Observed Empirical Mean under various Demographic History Scenarios[1] | | | | | | $\frac{E(t_i)}{E(t_j)}$ |
|---|---|---|---|---|---|---|---|
| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 | Scenario 6 | |
| $\hat{t}_2/\hat{t}_5$ | 11.56 | 4.15 | 4.50 | 18.12 | 20.54 | 8.70 | 10.00 |
| $\hat{t}_3/\hat{t}_5$ | 4.81 | 2.99 | 3.13 | 4.74 | 5.86 | 3.40 | 3.33 |
| $\hat{t}_4/\hat{t}_5$ | 2.83 | 2.46 | 2.70 | 2.57 | 2.46 | 2.73 | 1.67 |
| $\hat{t}_2/\hat{t}_4$ | 9.46 | 3.78 | 3.68 | 17.04 | 15.81 | 8.43 | 6.00 |
| $\hat{t}_3/\hat{t}_4$ | 3.51 | 2.80 | 2.61 | 3.52 | 4.25 | 2.77 | 2.00 |
| $\hat{t}_2/\hat{t}_3$ | 6.92 | 4.23 | 3.66 | 12.36 | 11.94 | 7.67 | 3.00 |

[1] Descriptions of the various demographic history scenarios are given in Table 5.1.

**Table 5.3**

Empirical Mean Values of $\hat{t}_i / \hat{t}_j$ under Different Demographic Scenarios when the Number of Segregating Sites = 30

| Statistic | Observed Empirical Mean under various Demographic History Scenarios[1] | | | | | | $\frac{E(t_i)}{E(t_j)}$ |
|---|---|---|---|---|---|---|---|
| | Scenario 1 | Scenario 2 | Scenario 3 | Scenario 4 | Scenario 5 | Scenario 6 | |
| $\hat{t}_2/\hat{t}_5$ | 11.64 | 4.06 | 3.72 | 12.94 | 16.44 | 7.05 | 10.00 |
| $\hat{t}_3/\hat{t}_5$ | 3.92 | 2.80 | 2.63 | 3.54 | 5.06 | 2.71 | 3.33 |
| $\hat{t}_4/\hat{t}_5$ | 3.27 | 2.40 | 2.55 | 3.29 | 2.00 | 2.29 | 1.67 |
| $\hat{t}_2/\hat{t}_4$ | 7.56 | 3.35 | 3.66 | 14.42 | 13.53 | 7.19 | 6.00 |
| $\hat{t}_3/\hat{t}_4$ | 3.54 | 2.96 | 2.43 | 3.10 | 4.02 | 2.52 | 2.00 |
| $\hat{t}_2/\hat{t}_3$ | 5.96 | 3.51 | 3.63 | 12.01 | 10.31 | 7.33 | 3.00 |

[1] Descriptions of the various demographic history scenarios are given in Table 5.1.

**Table 5.4**
Results of the 2-sample Kolmogorov-Smirnov Test of Equality of Distributions under various Demographic Scenarios when the Number of Segregating Sites = 50

| Statistic | Distributions under Scenarios[1] Compared | Value of the KS-statistic | p-value |
|---|---|---|---|
| $\hat{t}_2/\hat{t}_5$ | Scenario1-Scenario2 | 0.3053 | 0 |
| | Scenario1-Scenario3 | 0.2986 | 0 |
| | Scenario1-Scenario4 | 0.3538 | 0 |
| | Scenario1-Scenario5 | 0.2388 | 0 |
| $\hat{t}_3/\hat{t}_5$ | Scenario1-Scenario2 | 0.1625 | 0 |
| | Scenario1-Scenario3 | 0.1732 | 0 |
| | Scenario1-Scenario4 | 0.1005 | 0.0005 |
| | Scenario1-Scenario5 | 0.1360 | 0.0001 |
| $\hat{t}_4/\hat{t}_5$ | Scenario1-Scenario2 | 0.0715 | 0.0585 |
| | Scenario1-Scenario3 | 0.0539 | 0.3208 |
| | Scenario1-Scenario4 | 0.0610 | 0.2783 |
| | Scenario1-Scenario5 | 0.0555 | 0.3020 |
| $\hat{t}_2/\hat{t}_4$ | Scenario1-Scenario2 | 0.3859 | 0 |
| | Scenario1-Scenario3 | 0.3695 | 0 |
| | Scenario1-Scenario4 | 0.2200 | 0 |
| | Scenario1-Scenario5 | 0.2137 | 0 |
| $\hat{t}_3/\hat{t}_4$ | Scenario1-Scenario2 | 0.1323 | 0 |
| | Scenario1-Scenario3 | 0.1345 | 0 |
| | Scenario1-Scenario4 | 0.0423 | 0.7341 |
| | Scenario1-Scenario5 | 0.0842 | 0.0497 |
| $\hat{t}_2/\hat{t}_3$ | Scenario1-Scenario2 | 0.1998 | 0 |
| | Scenario1-Scenario3 | 0.1798 | 0 |
| | Scenario1-Scenario4 | 0.1902 | 0 |
| | Scenario1-Scenario5 | 0.1871 | 0 |

[1] Descriptions of the various demographic history scenarios are given in Table 5.1.

**Table 5.5**
Results of the 2-sample Kolmogorov-Smirnov Test of Equality of Distributions under various Demographic Scenarios when the Number of Segregating Sites = 40

| Statistic | Distributions under Scenarios[1] Compared | Value of the KS-statistic | p-value |
|---|---|---|---|
| $\hat{t}_2/\hat{t}_5$ | Scenario1-Scenario2 | 0.3769 | 0 |
| | Scenario1-Scenario3 | 0.3656 | 0 |
| | Scenario1-Scenario4 | 0.1669 | 0 |
| | Scenario1-Scenario5 | 0.2265 | 0 |
| $\hat{t}_3/\hat{t}_5$ | Scenario1-Scenario2 | 0.1751 | 0 |
| | Scenario1-Scenario3 | 0.1479 | 0 |
| | Scenario1-Scenario4 | 0.0438 | 0.7135 |
| | Scenario1-Scenario5 | 0.0830 | 0.0880 |
| $\hat{t}_4/\hat{t}_5$ | Scenario1-Scenario2 | 0.0765 | 0.0580 |
| | Scenario1-Scenario3 | 0.0543 | 0.3227 |
| | Scenario1-Scenario4 | 0.0690 | 0.2783 |
| | Scenario1-Scenario5 | 0.0555 | 0.6020 |
| $\hat{t}_2/\hat{t}_4$ | Scenario1-Scenario2 | 0.3195 | 0 |
| | Scenario1-Scenario3 | 0.3395 | 0 |
| | Scenario1-Scenario4 | 0.1900 | 0 |
| | Scenario1-Scenario5 | 0.2137 | 0 |
| $\hat{t}_3/\hat{t}_4$ | Scenario1-Scenario2 | 0.1132 | 0.0003 |
| | Scenario1-Scenario3 | 0.1329 | 0 |
| | Scenario1-Scenario4 | 0.0359 | 0.8451 |
| | Scenario1-Scenario5 | 0.0913 | 0.0133 |
| $\hat{t}_2/\hat{t}_3$ | Scenario1-Scenario2 | 0.1983 | 0 |
| | Scenario1-Scenario3 | 0.1778 | 0 |
| | Scenario1-Scenario4 | 0.1502 | 0 |
| | Scenario1-Scenario5 | 0.1671 | 0 |

[1] Descriptions of the various demographic history scenarios are given in Table 5.1.

**Table 5.6**
Results of the 2-sample Kolmogorov-Smirnov Test of Equality of Distributions under various Demographic Scenarios when the Number of Segregating Sites = 30

| Statistic | Distributions under Scenarios[1] Compared | Value of the KS-statistic | p-value |
|---|---|---|---|
| $\hat{t}_2/\hat{t}_5$ | Scenario1-Scenario2 | 0.2804 | 0 |
| | Scenario1-Scenario3 | 0.3208 | 0 |
| | Scenario1-Scenario4 | 0.1499 | 0 |
| | Scenario1-Scenario5 | 0.2896 | 0 |
| $\hat{t}_3/\hat{t}_5$ | Scenario1-Scenario2 | 0.1372 | 0 |
| | Scenario1-Scenario3 | 0.1819 | 0 |
| | Scenario1-Scenario4 | 0.0517 | 0.5690 |
| | Scenario1-Scenario5 | 0.1112 | 0.0102 |
| $\hat{t}_4/\hat{t}_5$ | Scenario1-Scenario2 | 0.0539 | 03921 |
| | Scenario1-Scenario3 | 0.0585 | 0.2948 |
| | Scenario1-Scenario4 | 0.0961 | 0.0701 |
| | Scenario1-Scenario5 | 0.0773 | 0.2822 |
| $\hat{t}_2/\hat{t}_4$ | Scenario1-Scenario2 | 0.2444 | 0 |
| | Scenario1-Scenario3 | 0.2382 | 0 |
| | Scenario1-Scenario4 | 0.2247 | 0 |
| | Scenario1-Scenario5 | 0.2386 | 0 |
| $\hat{t}_3/\hat{t}_4$ | Scenario1-Scenario2 | 0.1323 | 0 |
| | Scenario1-Scenario3 | 0.1345 | 0 |
| | Scenario1-Scenario4 | 0.0423 | 0.7341 |
| | Scenario1-Scenario5 | 0.0842 | 0.0497 |
| $\hat{t}_2/\hat{t}_3$ | Scenario1-Scenario2 | 0.1546 | 0 |
| | Scenario1-Scenario3 | 0.1829 | 0 |
| | Scenario1-Scenario4 | 0.2229 | 0 |
| | Scenario1-Scenario5 | 0.1562 | 0 |

[1] Descriptions of the various demographic history scenarios are given in Table 5.1.

as $\hat{t}_4 / \hat{t}_5$. Since the distributions were so widely different, these statistics can be used to infer and test contrasting demographic-history hypotheses of populations. For example, a small value of $t_2/t_5$ is more probable for a population that has a history of expansion, while we are more likely to get a large value of this ratio if the population underwent a bottleneck.

In order to use $t_i/t_j$ as a statistic to test the null model, we have to determine its sampling distribution. As we have seen earlier, the empirical distribution of $\hat{t}_i / \hat{t}_j$ is not significantly different from the empirical distribution of $t_i/t_j$. We can, therefore, determine the critical values of the statistic from the known distribution of $t_i/t_j$. Otherwise, one can obtain the critical values from a distribution that is close to the distribution of the statistic. For example, from the empirical distribution that can be obtained if independent samples can be generated under the null model. The coalescent program (Hudson 2002) generates samples when either the value of $\theta = 4N_e\mu$ or $S_n$ = *total number of segregating sites in the sample* is given. If the parameter is unknown, one approach is to obtain the critical values for a number of values of the parameter and to take either the maximum or the minimum of these critical values as the critical value of the test (Fu and Li 1993; Simonsen et al. 1995). Since the true values of the population parameter $\theta$ and its estimates vary greatly under different demographic scenarios (Tajima 1989), but $S_n$ for a specific data set is fixed, we suggest that for coalescent simulations to obtain empirical critical values, as suggested above, $S_n$ as observed in the data set under analysis be used.

## Summary

In Chapter 4, we have provided statistical evidence that even the estimate of the time to the most recent common ancestor (TMRCA) can be poor. Additionally, past demographic histories, such as whether the population size has remained constant or whether the population has passed through a bottleneck, affect the phylogenetic relationships among DNA sequences, particularly branch lengths. Even to estimate TMRCA, prior knowledge, or minimally some assumptions, of a population parameter $\theta=4N\mu$ ($\mu$=mutation rate/site/generation) is required, which is often unknown. In this Chapter we have proposed a statistical method to efficiently estimate *relative* branch lengths, which is often sufficient for evolutionary inferences. This method does not require prior knowledge or make any assumptions on $\theta$. In a sample of $n$ haploid DNA sequences, $E(k) = 2\,\mu\,B\,t$, where $k$ is the vector of the number of nucleotide mismatches between pairs of sequences, $B$ is a matrix of zeroes and 1s, and $t$ is the vector of divergence times between pairs of sequences. Using quadratic programming, we have obtained the least-squares estimators of the *relative* times of divergence, under the constraint $t \geq 0$, which are independent of $\mu$ and hence do not require prior knowledge of $\mu$. These estimates are sensitive to the demographic history of a population in a predictable way. We have studied the properties of these estimators using coalescent simulation techniques and have proposed a statistical method for drawing inferences on demographic history.

## Chapter 5: References


Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations. Genetics 141: 743-753

Chakraborty R (1977) Estimation of time of divergence from phylogenetic studies. Can J Genet Cytol 19: 217-223

Ewens WJ (1979) *Mathematical population genetics*. Springer, New York

Fitch WM, Margoliash E (1967) Construction of Phylogenetic Trees. Science 155: 279-284

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693-709

Goodman M, Barnabas J, Matsuda G, Moore GW (1971) Molecular evolution in the descent of man. Nature 233: 604-613

Hadley G (1964) *Nonlinear and dynamic programming*. Addison-Wesley Publishing Company, INC

Holmquist R (1972) Theoretical foundations for a quantitative approach to phylogenetics. Part I: DNA. J. Molec Evol 1: 115-133

Hudson R R (1990) Gene genealogies and the coalescent process. In Futuyma D and Antonovics J, editors, *Oxford Surveys in Evolutionary Biology* 7: 1-43. Oxford University Press, Oxford

Hudson R R (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338

Neumaier A (1998) MINQ - General Definite and Bound Constrained Indefinite Quadratic Programming (Software: http://www.mat.univie.ac.at/~neum/software/minq)

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J Mol Evol.19:153-70.

Nordborg M (2001) Coalescent Theory. In Balding D J, Bishop M J and Cannings C, editors, *Handbook of Statistical Genetics*: 179-212. John Wiley and Sons, Chichester, UK

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425

Simonsen KL, Churchill G, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595

Tateno Y, Nei M, Tajima F (1982) Accuracy of estimated phylogenetic trees from molecular data: I. Distantly related species. J Mol Evol 18: 387-404

Templeton AR (1993) The "eve" hypothesis: a genetic critique and reanalysis. Amer Anthrop 95: 51-72