UNIVERSITY OF TARTU
FACULTY OF BIOLOGY AND GEOGRAPHY, INSTITUTE OF MOLECULAR AND CELL BIOLOGY, DEPARTMENT OF EVOLUTIONARY BIOLOGY

Monika Karmin

HUMAN MITOCHONDRIAL DNA HAPLOGROUP R IN INDIA: DISSECTING THE PHYLOGENETIC TREE OF SOUTH ASIAN-SPECIFIC LINEAGES

M.Sc. Thesis

Supervisor: *Ph.D.* Ene Metspalu,

Tartu
2005

Table of contents

Abbreviations

(Y)BP	(years) before present
rCRS	revised Cambridge Reference Sequence
HVSI	the first hypervariable segment of the control region of mitochondrial genome
HVSII	the second hypervariable segment of the control region of mitochondrial genome
mtDNA	mitochondrial DNA
np	nucleotide position
RFLP	Restriction Fragment Length polymorphism

Definition of basic terms used in the thesis

haplotype	a sequence type that comprises all identical sequences
haplogroup	a group of haplotypes that share a common ancestor defined by an array of synapomorphic substitutions
lineage	any array of characters/mutations shared by more than one haplotype
coalescence	coalescence time calculated to the founder that displays star-like phylogeny

Introduction

Variability of human mitochondrial DNA has provided valuable data about the genetic past of our species**.** Analyses of the frequency, variation and distribution of mitochondrial DNA haplogroups have been used to evaluate current models concerning the process of colonization of the world. By now, there is already a substantial amount of mitochondrial evidence for an African exodus of humans, peopling of Eurasia, Australia and the Americas, as well as of the Pacific**.** South Asia lies on the way of earliest dispersals from Africa and is therefore a valuable well of knowledge on early human migrations. This thesis concentrates on the mitochondrial DNA haplogroup R sub-clusters in India. Full sequencing of mitochondrial genomes provides the best resolution for underlying phylogeny; therefore the task of fully sequencing 17 mitochondrial DNA molecules was undertaken. Better resolution of the phylogenetic tree provides a tool for further research for making more precise estimates of expansions of human populations and also their relations to each other in India and in Eurasia in general.

Part I: Literature overview

Characteristics of mtDNA

Mitochondrion is a cytoplasmic organelle of an eucaryotic cell which has a semi-autonomic genome. Human mitochondrial DNA (mtDNA) has about 16,6 kilo base pares (kbp) (Anderson *et al.* 1981; Andrews *et al.* 1999) forming a circular molecule. MtDNA codes 13 polypeptide genes for the oxidative phosphorylation complexes situated in the internal membrane of the mitochondrion. MtDNA codes for 12S and 16S rRNAs and also 22 tRNAs needed for mitochondrial protein synthesis (Wallace *et al.* 1995; Wallace *et al.* 1999). In mitochondria an alternative genetic code is used. MtDNA is built up very economically, the coding region is almost contiguous, the longest non-coding region of molecule spans between nucleotide positions 16024 and 575, and is called "control region", where also hypervariable regions HVS-I and HVS-II are situated (Taanman 1999). Human mtDNA is inherited maternally and it does not recombine, therefore acts as a single haploid locus (Giles *et al.* 1980; Chen *et al.* 1995a; Ingman and Gyllensten 2001; Herrnstadt *et al.* 2002; Sutovsky *et al.* 2004). In most postmitotic human cells there are about 1000 – 100 000 mtDNA molecules (Lightowlers *et al.* 1997) therefore making it easily assessable with tools of molecular biology.

MtDNA diverges at the rate of 2 - 4% per site per million years (Cann *et al.* 1987; Torroni *et al.* 1994), which is on the average tenfold faster than the rate in the nuclear genome. The high substitution rate has been attributed to the lack of proofreading activity in mitochondrial DNA polymerase and because of a high concentration of oxidative radicals inside mitochondria**.**

There have been many attempts to characterize the relative mutation rates in mtDNA, especially the two HVS regions (Hasegawa *et al.* 1993; Wakeley 1993; Malyarchuk *et al.* 2002). In HVS-I and HVS-II a few sites ("hotspots") exhibit very high mutation rates while most sites have a range of low substitution rates. The hotspot mutations occur frequently in many different phylogenetic context (Hasegawa *et al.* 1993; Wakeley 1993; Ingman *et al.* 2000; Stoneking 2000). The mutation rate in HVS-I has been estimated for the region from nucleotide positions (np) 16090-16365 and is 20180 years per transition (Forster *et al.* 1996). The fast substitution rate makes it possible to distinguish relatively

recently diverged populations, which is another reason why mtDNA is extensively used in population genetics studies. Historically the molecule has been assessed either by restriction fragment length polymorphisms and/or direct sequencing of HVS-I region. Phylogenies – graphic reconstruction of descendance between individual lineages (individuals), are reconstructed on the basis of mtDNA mutations that accumulate over time. The cladistic haplogroup nomenclature combined all available information (Torroni 1996; Richards 1998; (Macaulay et al. 1999). With the growing number of published mitochondrial whole genome sequences the haplogroup nomenclature based on diagnostic RFLP markers and HVS-I sequences has been confirmed (Finnilä *et al.* 2001; Herrnstadt *et al.* 2002; Kong *et al.* 2003). Also characterisation of relative mutation rates in the coding region is now feasible. The heterogeneity of mutation rate also applies to coding region (Herrnstadt *et al.* 2002). Coding region mutation rate - 5140 years per base substitution (every mutation other than insertion or deletion) has been calibrated on the basis of assumed human-chimp split 6,5 million years ago. The "coding region" for this calculation spans from nucleotide position 577 to 16023 (Mishmar *et al.* 2003).

MtDNA global phylogeny

Human maternal lineages have been classified into haplogroups (monophyletic clades). Basal haplogroups display continent specificity (Wallace *et al.* 1999; Ingman *et al.* 2000; Richards *et al.* 2000; Maca-Meyer *et al.* 2001; Herrnstadt *et al.* 2002). African populations are characterised by the super-haplogroups, L0 – L6. The likely root of human mtDNA tree is between haplogroups L0 and L1 dividing the phylogenetic tree into two basic clades: L0 and the rest. (Bandelt *et al.* 1995; Chen *et al.* 1995b; Graven *et al.* 1995; Watson *et al.* 1997; Alves-Silva *et al.* 2000; Chen *et al.* 2000; Torroni *et al.* 2001b; Salas *et al.* 2002; Kivisild *et al.* 2004; Salas *et al.* 2004). It seems that only L3 radiated out of Africa, in the form of haplogroups M and N, about 60,000 ybp, giving rise to the extant Eurasian variation (Quintana-Murci *et al.* 1999; Wallace *et al.* 1999). Most western Eurasians are characterized by clades within haplogroup N (Torroni *et al.* 1996; Macaulay *et al.* 1999; Richards *et al.* 2000; Finnilä *et al.* 2001; Herrnstadt *et al.* 2002; Palanichamy *et al.* 2004), whereas N and M contributed almost equally to the current eastern Eurasian mtDNA pool (Stoneking *et al.* 1990; Ballinger *et al.* 1992; Torroni *et al.* 1993; Horai *et al.*

1996; Kolman and Bermingham 1997; Comas *et al.* 1998; Starikovskaya *et al.* 1998; Derbeneva *et al.* 2002b; Kivisild *et al.* 2002; Schurr and Wallace 2002; Yao *et al.* 2002).

Western Eurasian mtDNA lineages converge into haplogroups HV, N1I, N2W, R1, R2, JT, UK and X. (Torroni *et al.* 1998; Macaulay *et al.* 1999; Richards *et al.* 2000; Torroni *et al.* 2001a; Reidla *et al.* 2003; Loogväli *et al.* 2004; Tambets *et al.* 2004).

Asian, Oceanian and Native American mtDNA lineages are stemming both from ancestral node N and M. East-Eurasian specific derivates of ancestral node R one daughter clade of N, are haplogroups B, P, S, R9, F, N5, R5 - R8, R30, R31. Haplogroups M1-M11, M18, M25, M30, M31, M32, D, G, Q derive from ancestral node M. (Kivisild *et al.* 2002; Kivisild *et al.* 2003b; Kong *et al.* 2003; Metspalu *et al.* 2004; Palanichamy *et al.* 2004; Rajkumar *et al.* 2005; Thangaraj *et al.* 2005). The skeleton of global phylogenetic tree is depicted in Figure 1.

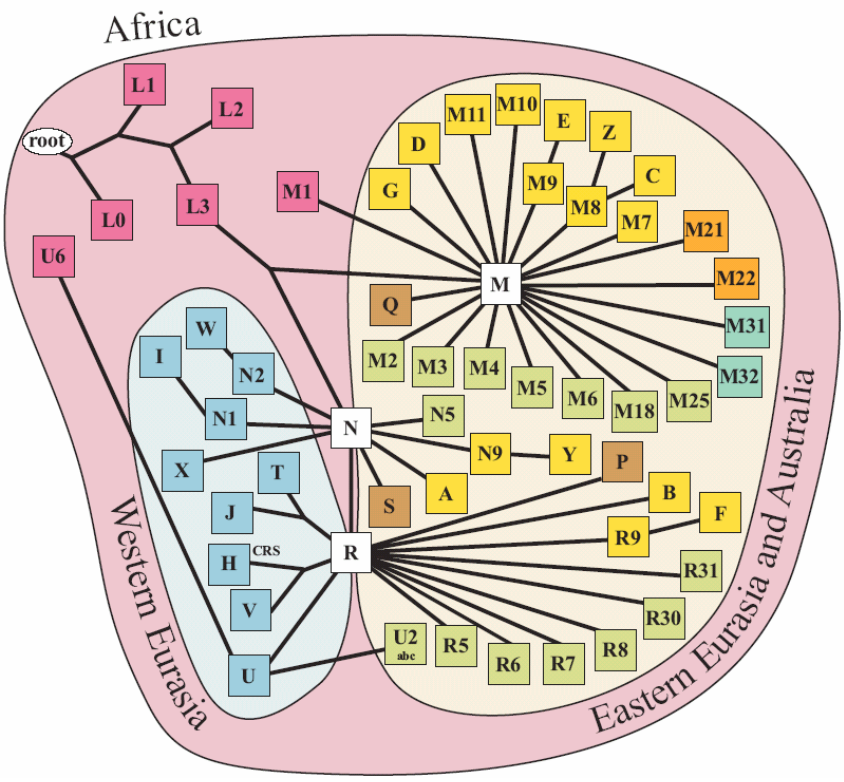


Figure 1. A skeleton of the global phylogenetic tree. Sub-branches of major African and Western Eurasian haplogroups are omitted. Colours: green – haplogroups specific for Indian subcontinent; yellow – Eastern Eurasian haplogroups; blue – Western Eurasin haplogroups; orange – Orang Asli haplogroups (Malaysian tribes); turquoise - haplogroups on Andaman Island; brown – Near Oceania; pink – African haplogroups. The tree is rooted by chimpanzee lineage (based on Ingman *et al* 2000). Data from: (Chen *et al*. 1995b; Macaulay *et al*. 1999; Chen *et al*. 2000; Bandelt *et al*. 2001; Finnilä *et al*. 2001; Maca-Meyer *et al*. 2001; Herrnstadt *et al*. 2002; Kivisild *et al*. 2002; Kong *et al*. 2003; Palanichamy *et al*. 2004; Macaulay *et al*. 2005; Thangaraj *et al*. 2005)

Out of Africa

Archaeological and fossil evidence suggests that modern humans originated in Africa more than 200 000 years before present (ybp) (Grun *et al.* 1998; McDougall *et al.* 2005). Current understanding found on genetic evidence is that modern humans arose about 150 000 years ago, possibly in East Africa, were human genetic diversity is particularly high (Cann *et al.* 1987; Chen *et al.* 1995b; Kivisild *et al.* 2004). The recent dispersal of modern humans out of Africa is now widely accepted, while more discussion is held over the routes taken across Eurasia. A northern route out of Africa using the corridor of Nile river over Sinai peninsula to Near East (Lahr and Foley 1994; Maca-Meyer *et al.* 2001; Tanaka *et al.* 2004).

The existence of an early southern route has been supported by analyses of mtDNA restriction enzyme data from New Guinea (Forster *et al.* 2001) and control region sequences from mainland India and the Andaman Islands (Endicott *et al.* 2003; Kivisild *et al.* 2003a; Metspalu *et al.* 2004; Macaulay *et al.* 2005). Similarly, archaeological research has proven the presence of modern humans in the Red Sea region as early as 125000 years ago (Walter *et al.* 2000). However, the published genetic data remain sufficiently ambiguous for some geneticists to reject the very existence of the southern route (Cordaux *et al.* 2003), first suggested already by Carl Sauer (Sauer 1962). The question of single versus multiple dispersals remains disputed still. Recent findings favour one single dispersal from Africa via southern route, through India and onward to southeast Asia and Australasia. There was an early offshoot, leading ultimately to the settlement of the Near East and Europe, but the main dispersal from India to Australia ~65 000 years ago was rapid, most likely taking only a few thousand years (Macaulay *et al.* 2005).

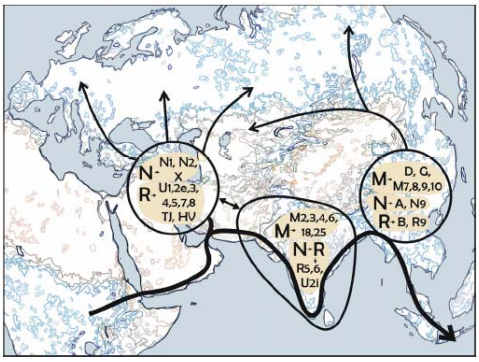


Figure 2. Two proposed routes out of Africa – the Northern and the Southern route (Metspalu 2004)

The southern route scenario puts India on the way of earliest migrations of anatomically modern humans. The lack of L3 lineages other than M and N in India (Roychoudhury *et al.* 2000; Kivisild *et al.* 2003a; Kivisild *et al.* 2003b) and among non-African mitochondria in general (Ingman and Gyllensten 2001; Herrnstadt *et al.* 2002; Kivisild *et al.* 2002) suggests that the earliest migration(s) of modern humans already carried these two mtDNA ancestors. The departure route over the horn of Africa, the southern route migration led people towards South and East Asia (Nei and Roychoudhury 1993; Quintana-Murci *et al.* 1999; Stringer 2000) The N branch had already given rise to its daughter clade R, which later, in eastern Asians, differentiated into clusters B and R9 (Kivisild *et al.* 2002) and in western Eurasia gave rise to haplogroups HV, TJ, and U (Macaulay *et al.* 1999). The deep coalescence times of some major M and R sub-clusters in the Indian subcontinent suggests that it was settled soon after the African exodus and that there has been no complete extinction or replacement of the initial settlers (Kivisild *et al.* 1999a; Kivisild *et al.* 1999b; Kivisild *et al.* 2000; Kivisild *et al.* 2003b; Metspalu *et al.* 2004).

India

India is geographically, culturally and anthropologically a highly heterogeneous region. The total population of India by 2001 census was 1,027,015,247 (http://www.censusindia.net/). This fact alone is a good reason for carrying out genetic studies on Indian populations. India has long been reckoned as the most stratified of all known societies in human history. The population is divided according to linguistic and religious affiliation. Within each religious and linguistic group there are several strata of cultural, social and biological differences. The caste system with its myriad forms of superordination and subordination is perhaps most known of these. In general the populations can be grouped into caste and tribal peoples. The tribal people, called officially the scheduled tribes, constitute about 7.8% of the total population of India and might represent relict populations (Singh 1997).

As a local proverb has it, "Every two miles the water changes, every four miles the speech". There are a total of 114 languages and 216 mother tongues by the Census of India 1991 (www.languageininidia.com). The languages in India belong to four major families: Indo-Aryan (a branch of the Indo-European family), Dravidian, Austro-Asiatic (Mundari branch), and Sino-Tibetan, with the overwhelming majority of the population speaking

languages belonging to the first two families. The geographical range of distribution of Austro-Asiatic, Indo-European and Sino-Tibetan speaker is extensive, India harbours only a fraction of the languages within these families. Dravidian languages are restricted largely to India; there are two outlying populations –Brahui in Baluchistan. In India the language families have their characteristic geographical distribution. Indo-Aryan speakers (80% of Indian population) are spread mainly across northern and central India. About 18% of Indian populace speak Dravidian languages. Most Dravidian speakers reside in south India; only a few isolated groups of Dravidian speakers remain in the central and eastern India (in Madhya Pradesh, Orissa and Bihar). Sino-Tibetan speakers live along the Himalayan fringe from Jammu and Kashmir to eastern Assam and they comprise about 1.3 % of the whole population of India. Languages of the Mundari branch of the Austro-Asiatic family are spoken by groups of tribal people from West Bengal through Bihar and Orissa and into Madhya Pradesh. These groups make up approximately 0.7% of the whole population.

Indian genetic variability studies based on classical markers

Indian peoples have been subjects to numerous anthropological and genetic studies (for a review see (Cavalli-Sforza *et al.* 1994; Papiha 1996). It has been uncertain weather the genetic diversity seen today reflects primarily their local long-term differentiation or is due to relatively recent migrations from abroad. First studies used "classical" genetic markers i.e. geographic mapping of allele frequencies. A review of the studies from 1970s to the date of the article has been provided by S. S. Papiha in 1996 (Papiha 1996). Most of the studies cover limited number of genetic systems, and only a few of them provide information on the genetic differentiation and population structure of some tribal, caste, religious, and urban groups. In conclusion, Papiha claims that tribal populations are in general well differentiated from the nontribal castes or communities. Genetic differentiation among nontribal communities and occupational castes is slight, but the subpopulations of each nontribal group of different provinces demonstrate considerable genetic diversity.

Indian maternal lineages

From the time when mtDNA became the one of the main tools in tracing human prehistoric movements, also Indian maternal lineages have been under study. India lies between West Eurasia and East Eurasia, which is mirrored in the genetic composition of Indian mtDNA-lineages. More than 60% of present-day Indian maternal lineages descend from (super)haplogroup M (Roychoudhury *et al.* 2000; Kivisild *et al.* 2003b; Metspalu *et al.* 2004), which split into Indian, eastern Asian, Papuan and Malaysian (Macaulay *et al.* 2005) and Australian subsets 40 000 – 60 000 years ago. The second major component of Indian mtDNA lineages stem from a N (and R derivate) haplogroup U, a complex mtDNA lineage cluster with an estimated age of 51,000–67,000 years (Kivisild *et al.* 1999b). There are also lineages stemming straight from the ancestral nodes N and R as well as R-derivates (besides U) – characteristic for western Eurasia and eastern Asia and M derivates characteristic for eastern Asia (Kivisild *et al.* 1999b; Roychoudhury *et al.* 2001; Metspalu *et al.* 2004).

Haplogroup M

The sub-branching of haplogroup M in India is profoundly different from that described so far for any other Asian locality. More than 60% of Indians have their maternal roots in Indian-specific branches of haplogroup M. Typical Mongoloid sub-clusters C, D, E and G are found at extremely low frequencies (Kivisild *et al.* 1999b). Indian specific M-subclades defined so far are - M2, M3a, M4a, M6, M18, M25, M30, M31 (the last two are at odds with the rest of published and also our labs unpublished data). There are yet others lineages, which can not be classified further from the basal node M* (Kivisild *et al.* 1999a; Metspalu *et al.* 2004; Rajkumar *et al.* 2005).

Nearly one tenth of the Indian haplogroup M mtDNAs fall into its major sub-clade **M2** (Kivisild *et al.* 2003a). M2 can be further subdivided into haplogroups M2a and M2b. The coalescence times for M2 and its major sub-clades are 50 000 – 70 000 ybp. Haplogroup M2 frequency increases towards the southern part of India, which makes it is significantly more frequent among Dravidic speakers than among the Indo-European speakers (Metspalu *et al.* 2004).

Sub-clades of **M3** and **M4** display a different geographic distribution. M4a is sparsely spread in most of India with no obvious geographical cline. The spread of M3a is concentrated into north-western India, suggesting the ancestral source region.

M6 is primarily found in the Indus Valley and on the western shores of the Bay of Bengal. Sub-clades of M6 – M6a, M6b are concentrated towards the southwest and the northeast. **M18** is spread at low frequencies across India, except for far north and the coast of Arabian Sea. **M25** is moderately common in Kerala and Maharashtra but rather infrequent elsewhere in India (Metspalu *et al.* 2004).

Haplogroups stemming from the node N

With the exception of the diverse set of largely Indian-specific R lineages, the most frequent mtDNA haplogroup in India that derives from the phylogenetic node N is haplogroup W (Kivisild *et al.* 1999a; Kivisild *et al.* 1999b). The frequency peak of haplogroup W is 5% in the north-western states – Gujarat, Punjab and Kashmir. Elsewhere in India its frequency is very low (from 0 to 0.9%) forming a significant spatial cline (Metspalu *et al.* 2004). Haplogroup N2 stems straight from the node N, but there its frequency distribution is not studied yet (Palanichamy *et al.* 2004).

Haplogroup R and its derivates in India

The most frequent sub-clade of R in India is haplogroup U, with Indian-specific variants of U2 (a, b, c) (Kivisild *et al.* 1999a; Quintana-Murci *et al.* 2004). Haplogroup U, mtDNA lineage cluster with an estimated age of 51,000–67,000 years, represents the most profound overlap between western-Eurasian and Indian mtDNA lineages. The frequency of haplogroup U in India reaches 15% among the caste and 8% among the tribal populations (Kivisild *et al.* 1999a; Kivisild *et al.* 2000; Metspalu *et al.* 2004).

The most common sub-cluster of U in India are Indian-specific clades of U2 (U2i, U2a, U2b and U2c), which do not show a distinct geographic cline and are present throughout India and coalesce with western-Eurasian U2 lineages 53,000 ± 4,000 ybb (Kivisild *et al.* 1999a; Quintana-Murci *et al.* 2004); Metspalu 2004).

Another subset of U, haplogroup U7, is also present in India (Kivisild *et al.* 1999a). The distribution of U7 ranges from India to Iran, and has been found also in some European and Siberian populations (Richards *et al.* 2000; Derbeneva *et al.* 2002a). The peak of U7 frequency is at 12 % in Gujarat and at 9 % in Punjab. For the whole of India its frequency is around 2 %. Expansion times and haplotype diversities for the Indian and the Near and the Middle Eastern (41000 ± 15800 and 41200 ± 14800 respectively) U7 mtDNA are similar. It is possible that this haplogroup started to expand from the territories of today's Gujarat and Iran. This is evident from the high frequencies of these haplogroups in Gujarat and Iran and the diverse haplotypes present in those localities (Metspalu *et al.* 2004).

Over two third of western Eurasian specific maternal lineages in the Indian mtDNA pool belong haplogroups HV, pre-HV, I, N, JT, U2e, W and X which are represented by low overall frequencies (Kivisild *et al.* 1999b; Metspalu *et al.* 2004).

Until the paper of Palanichamy and colleagues was published in the fall 2004, many Indian lineages could not be distinguished from the ancestral node R* (Kivisild *et al.* 1999b; Basu *et al.* 2003; Kivisild *et al.* 2003a; Metspalu *et al.* 2004; Quintana-Murci *et al.* 2004). In the aforementioned paper there were altogether 27 mitochondrial genomes belonging to haplogroup R fully sequenced. As a consequence four new sub-haplogroups were defined (R7, R8, R30, R31) and already known haplogroup definitions (R5 and R6) were specified.

According to the resolved phylogeny of haplogroup R* seven coding region mutations (1442, 6248, 9051, 9110, 10289, 13105, 13830) and four control-region mutations (16260, 16261, 16319) define haplogroup R7.

Haplogroup R8 is recognisable by five specific mutations (2755, 3384, 7759, 9449, 13215). Haplogroup R30 is tentatively defined by 8589 transition and haplogroup R31 with a transition at nucleotide position 15884. At the same time it is believed that this classification needs further refinement, since recurrent mutations have been observed in both of these sites (Metspalu *et al.* 2004; Palanichamy *et al.* 2004).

Refinements of known haplogroups included haplogroup R5, which in addition of having mutations at 8594 (corresponding to -8592*Mbo*I) and 16304 (Quintana-Murci *et al.* 2004), is further characterized by another three mutations, at 10754, 14544, and 16524. It is not clear at this point whether the 16266 transition defines R5 as a whole or just the main sub-branch of R5, since 16266 seems to be prone to back mutation in this haplogroup. Similar caveat has been observed for hypervariable site 16129 in haplogroup R6 (Palanichamy 2004). R6 has been defined by -12282*Alu*I and transitions at nps 16129 and 16362 (Quintana-Murci *et al.* 2004). For published phylogeny of haplogroup R lineages in India see Figure 3.

Haplogroup R2 is defined by coding region motif +4216*Nla*III, +4769*Alu*I, -14304*Alu*I and control region mutation 16071 (Quintana-Murci *et al.* 2004) There is also one published R2 full genome sequence (Palanichamy *et al.* 2004). R2 is present in India at low frequencies.

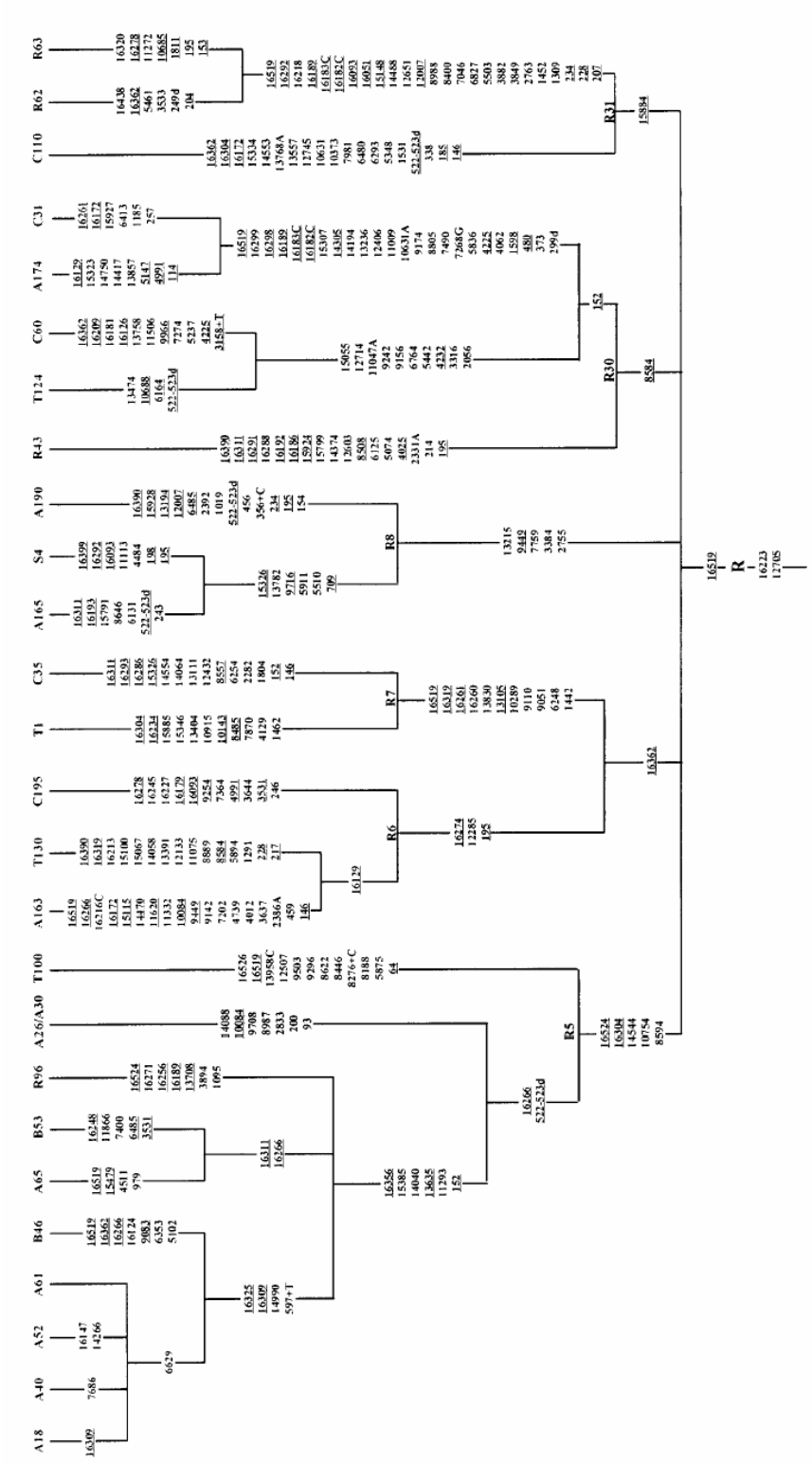


Figure 3. Haplogroup R phylogeny (Palanichamy *et al.* 2004).

Phylogeography of R lineages in India.

Haplogroup R2 is present at low frequencies in Near and Middle East, India and Pakistan where the coalescence age of the haplogroup is about forty thousand years (Metspalu *et al.* 2004; Quintana-Murci *et al.* 2004). The spread of this haplogroup in Europe is restricted to a few populations in the Volga basin where it is represented by nodal haplotypes and by a region-specific sub-clade with the coalescence estimate of 11400 ± 9000 ybp (Metspalu *et al.* 2004).

R5 is the second most frequent sub-clade of R in India after haplogroup U. The coalescence age estimate for R5 is 50 000 to 70 000 ybp whereas individual sub-clades of R5 have shown expansion times ranging from 20000 ypb to 50 000 ybp. This diverse and ancient haplogroup is present over most of India. R5 is spread across language boundaries, but is absent among the Austro-Asiatic tribal groups. R5 is more frequent among caste (2,5 – 4,5%) than among the tribal populations (0,7 – 1,9%) (Metspalu *et al.* 2004).

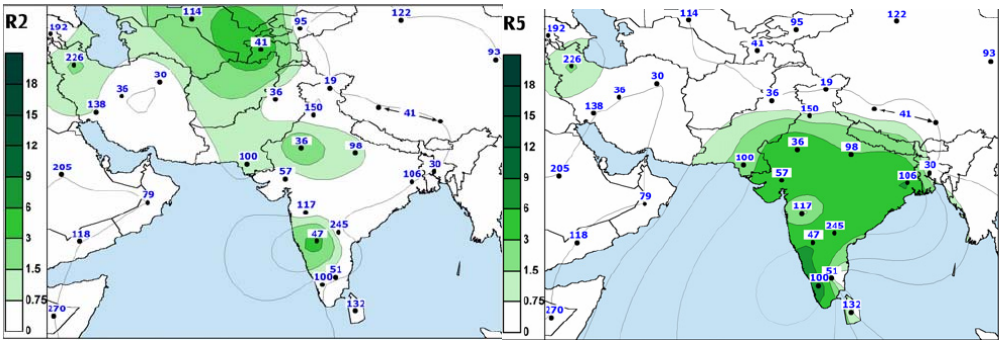


Figure 4. Phylogeography of haplogroups R2 and R5 (Metspalu *et al.* 2004)

Haplogroup R6 is present almost all over India, with equal frequencies form 1,5 to 3%. R6 has been found also in neighbouring Pakistani population (Metspalu *et al.* 2004; Quintana-Murci *et al.* 2004)

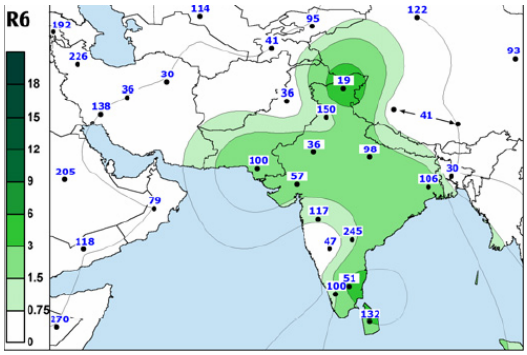


Figure 5. The spread of haplogroup R6 in India (Metspalu *et al.* 2004)

Various scenarios, including multiple expansions out of Africa could explain the genetic patterns in current Indian populations. Space, time or both could have separated these from each other. Parsimony criterion supports a single early migration, which brought ancestral lineages M and N (with already descended R) to South Asia. This early migration has shaped the extant phylogeography in Eurasia (Kivisild *et al.* 2003a).

Indian specific sub-clades of haplogroup U2 (Kivisild *et al.* 1999a) Quintana-Murci 2004), M2 and R5 show extremely deep coalescence times (50 000 – 70 000 ypb). Together they constitute nearly 15% of the Indian mtDNAs. These haplogroups are virtually absent elsewhere in Eurasia (Kivisild *et al.* 1999a; Kivisild *et al.* 2003a; Metspalu *et al.* 2004) with the exception of a few occasional R5 or U2 abc samples from Cental Asia or Iran (Metspalu *et al.* 2004; Quintana-Murci *et al.* 2004). The time depth for the arrival of the predicted founder R haplotype in India is 64 200 ±6300 ybp (Palanichamy *et al.* 2004).

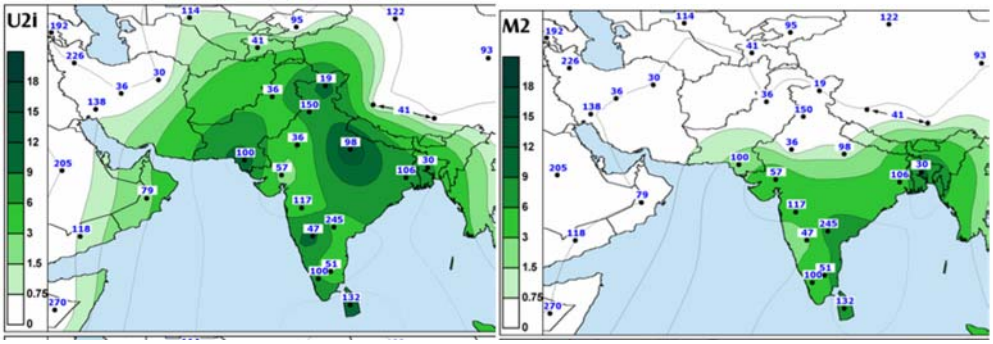


Figure 6. The spread of most ancient haplogroups U2i and M2 (for R5 see Fig 4.) (Metspalu *et al.* 2004)

Haplogroups U7, R2 and W show a time depth about 40 000 years before present and their phylogeography forms a genetic continuum spanning from the Near East into India extending also north into Central Asia (Metspalu *et al.* 2004).

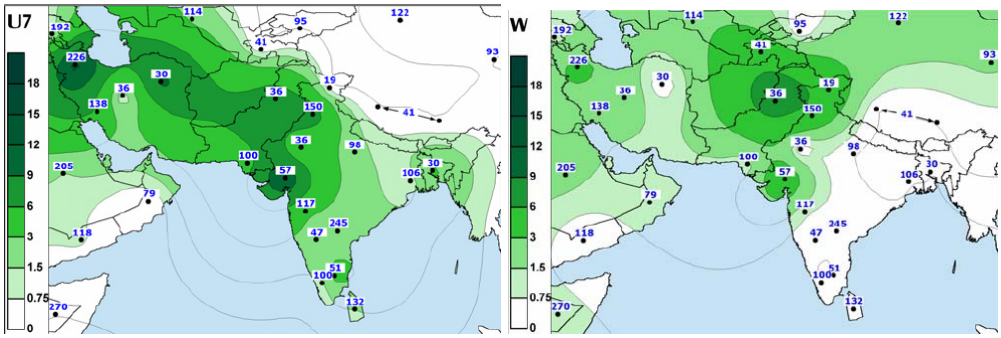


Figure 7. Genetic continuum of haplogroups W and U7 from Near East to Central Asia and India (for R2 spreading pattern see Figure 4) (Metspalu *et al.* 2004)

Indian-specific mtDNA haplogroups M3a, M4a, M6, M25, R6 reveal coalescence estimates falling largely between 20000 and 30000 ybp. These estimates overlap with those of many West Eurasian-specific (e.g. H, HV, preHV, U3, U, K, X (Richards *et al.* 2000; Reidla *et al.* 2003; Metspalu *et al.* 2004) and East Eurasian specific (A, F2, D4, M7c1, m7a1, M8a (Kivisild *et al.* 2002; Yao and Zhang 2002) mtDNA clades, suggesting a rather synchronic worldwide demographic expansion event in the late Pleistocene, during an interglacial period preceding the LGM (Metspalu *et al.* 2004).

The spread of haplogroups with deep coalescence ages (M2, U2i and R5) can be associated with the initial peopling of South Asia. The overlapping geographic distributions and coalescence times suggest a deep autochthonous history of haplogroups U7, W and R2 in India. The sharp decline in the haplogroup M frequency between India and Iran marks the western border of this haplogroup. Similarly distinct borders are observed in the distribution of Indian-specific mtDNA haplogroups to the east and to the north of the subcontinent. This has led to the proposal that the initial South Asian mtDNA pool was established upon the initial peopling of this region and has not been replaced but rather reshaped locally by major demographic episodes in the past (Metspalu *et al.* 2004).

Socio-linguistic and genetic variability in India

Studies based on mtDNA have shown that the basic clustering of lineages in India is not language or caste-specific, although a low number of shared haplotypes indicates that recent gene flow across linguistic and caste borders has been limited (Bamshad *et al.* 2001; Roychoudhury *et al.* 2001; Kivisild *et al.* 2003b; Metspalu *et al.* 2004).

India congregates four linguistic domains, which all have their characteristic geographical range. The majority of studies based on mtDNA variation have provided evidence that linguistic groups of India do not represent genetically homogeneous units and are not traceable to different immigration waves from distinct sources (Kivisild *et al.* 1999a; Kivisild *et al.* 1999b; Bamshad *et al.* 2001)

There have been speculations that populations affiliated to the Austro-Asiatic linguistic family were probably the earliest to settle in the subcontinent either from North East (Diamond 1988; Ballinger *et al.* 1992; Gadgil 1997)or from North West (Chu *et al.* 1998; Su *et al.* 1999; Majumder 2001a). The criticism has pointed out that Austro-Asiatic groups in India represent more than 30 endogamous tribal populations, but only few were included in the studies, disregarding the implicit genetic/linguistic heterogeneity among these tribes. Also small sample sizes were used. Studies, which have shown differences in haplotype sharing among tribals, have concluded that Austro-Asiatic tribal groups are the oldest inhabitants of India on the basis of nucleotide diversity (Majumder 2001a; Roychoudhury *et al.* 2001; Basu *et al.* 2003). Weak points of these works were the incorrect assignment of haplogroups (based only on HVS-I region). The results for nucleotide diversity calculations can be misleading because linguistic groups do not form specific lineage clusters on Indian mtDNA tree. Different linguistic groups are spread equally over the phylogenetic tree and the differences in haplogroup frequencies are shaped by geography. Language families in India are much younger than the mtDNA lineages their speakers harbour, therefore it is rather speculative to infer to some linguistic families as more autochthonous to India (Metspalu *et al.* 2004).

A good example of social processes modulating evolutionary processes is provided by Hindu caste system, which governs the mating practices of nearly one-sixth of the world's population (Bamshad *et al.* 1998; Majumder 2001b). Different social rank between castes

corresponds to mtDNA distances (Bamshad *et al.* 1998; Bamshad *et al.* 2001). No haplotypes are shared exclusively between upper and lower castes. This suggests that haplotype sharing between castes is limited by social rank. The genetic distance between upper and lower castes (0.00045) is 1.5 times greater than between the upper and middle (0.00024) or middle and lower castes (0.00030). Thus, mtDNA distances between castes of different status are stratified according to social rank. This is consistent with higher levels of female gene flow between more closely ranked castes. Paternal lineages do not display this kind of correlations, which is in concordance with ethnographic data. Genetic stratification of the Hindu caste system is driven by the social mobility of women (Bamshad *et al.* 1998). Phenomena like the upward social mobility of caste women could have introduced some tribal genes to the castes more recently. Given the relatively low proportion of the tribal population size today, recent unidirectional gene flow can be assumed to be a minor modifying force in the formation of the genetic profile of the caste population (Bamshad *et al.* 1998). When the genetic patterns in tribal populations from southern India were compared to those of caste populations from different regions in the subcontinent, the sharing of some deep lineages was revealed. Among them was haplogroup M2 that has deep coalescence time and high diversity in India signalling long autochthonous development. It was possible to trace the genetic heritage of earliest settlers of Indian sub-continent both to tribal and caste populations, thus the idea of recent large-scale replacement of populations in India (Indo-Arian invasion theory) did not find support. Caste populations have higher share of western Eurasian-specific lineages because they have been subjects to stronger influence from West Eurasian gene pool, but share deep Indian lineages with tribals (Kivisild *et al.* 2003a).

II Experimental work

The aim of the present study

In order to infer the prehistory of populations correctly a phylogenetic tree without ambiguities is essential. It is important to be able to separate deep-rooting clusters from the shallow ones to dissect essentially old lineages from younger ones. The maximum resolution of mtDNA phylogenetic tree is achieved when the information about the nucleotide contents of the whole genome is known. Only then the tree containing all the information can be constructed. Once the topology of a phylogenetic tree is established firmly, only truly informative nucleotide changes can be studied in a wider variety of genomes in order to integrate them correctly into the phylogeny. This provides a wider view of the frequency and spreading patterns of given lineages.

Complete mtDNA information is highly relevant also to medical studies of mitochondrial diseases. The basal outline of the total mtDNA phylogeny in a region is indispensable to perform systematic studies of major mitochondrial diseases.

The main goal of the present study was to further clarify the topology of mtDNA phylogenetic tree of Indian (and South Asian) specific lineages by fully sequencing mitochondrial genomes. As a second step the informative sites from complete sequence phylogeny were studied in a larger sample set to clarify the geographic spread of these haplogroups.

Materials and Methods

Sample Selection for Whole Genome Sequencing

Initially 17 samples from India (Table 1) were selected, which could not be further classified from the ancestral node R* (according to their HVS-I sequence). The samples were collected during 1970s under the supervision of Dr. Surinder S. Papiha and Dr. Sabjit Mastana.

Table 1. Studied Individuals

State	social status	caste/ religion	language group	Population	ID	HG	HVS1
Uttar Pradesh	caste	N/A	Indo-European	Uttar Pradesh	606	R2	71
Rajastan	caste	N/A	Indo-European	Rajastan	487	R2	71-293
Sri Lanka	caste	Muslim	Dravidic	Moor	38	R5	266-304-309-325-356
West Bengal	caste	Brahmin	Indo-European	Bengal	46	R5	266-304
Gujarat	caste	N/A	Indo-European	Gujarat	35	R5	266-304-309-325-356
Andhra Pradesh	tribe		Dravidic	Koya	5	R6	129-266-318-320-362
Andhra Pradesh	tribe		Dravidic	Koya	18	R7	260-261-311-319-362
Maharashtra	caste	Brahmin	Indo-European	Konkanastha Brahmin	41	R8	CRS
Andhra Pradesh	tribe		Dravidic	Koya	30	R8	324
Andhra Pradesh	tribe		Dravidic	Koya	37	R8	324
Andhra Pradesh	tribe		Dravidic	Koya	74	R8	CRS
Andhra Pradesh	tribe		Indo-European	Lambadi	10	R8	292
Sri Lanka	caste	Goyigama	Indo-European	Sinhalese	49	R30	172-278
Punjab	caste	Jat Sikh	Indo-European	Punjab	47	R30	92-189-298-299
Maharashtra	caste	Brahmin	Indo-European	Konkanastha Brahmin	23	R30	286-292
Rajastan	caste	Kshatriya	Indo-European	Rajput	25	R31	172-304-362
Rajastan	caste	Kshatriya	Indo-European	Rajput	48	R31	172-304-362

In haplogroup frequencies and distribution calculations the database of Indian genetic variability in TU ICMB Department of Evolutionary Biology was used. The total number of samples was 2565 for haplogroups R8, R30 and R31. For frequency and distribution analyses of haplogroup R7 Kumarasamy Thangaraj kindly provided the unpublished data from Centre for Cellular and Molecular Biology, Hyderabad, India, so the total number of samples was 8216.

PCR amplification of the mtDNA Genome

DNA amplification was performed on 10 – 20 ng of template DNA and was carried out with the thermocycler “Biometra UNO II” usually in total volume of 25 μl.

1/10	Tartrazine Buffer (750 mM Tris-HCl, pH 8.8, 200 mM $(NH_4)_2SO_4$, 0.1% Tween 20; 5% Ficoll 400.)
2,5 mM	$MgCl_2$ (25 mM $MgCl_2$)
1 mM	dNTP mix (dATP, dCTP, dGTP, dTTP; 10 mM)
0.2 –0.35U	Taq DNA polymerase (FIREPol) from Solis Biodyne and/or LongExtention Taq2 U/µl
~0,2 pM	Forward primer (10 pmol/µl)
~0,2 Pm	Reverse primer 10 pmol/µl
1-2 µl	DNA sample
	Deionized water

Amplification cycle for PCR:

Denaturation	94°C	2 min	
Denaturation	94°C	25 s	42-44 cycles
Primer annealing	52-61°C	20 s	
DNA Syntheses	72° C	1.5 – 2 min	
Final Syntheses	72° C	3 min	

Number of cycles and annealing temperature depended on primer specifity and mtDNA quality. All the primers used for attaining PCR products and sequencing are listed in the table1 in supplementary materials. Due to heterogeneous quality of the used DNA many different primers were used both for PCR products and sequencing to fill the gaps in sequences.

An aliquot of 3 µl of PCR product was fractionated by gel electorphoreses in a 2% agarose gel containing 0,5 µg/mL of ethidium bormide to assess the purity and size of the DNA fragments.

All PCR products were purified with Shrimp Alkaline Phosphatase (SAP)/Exonuclease I (ExoI) (Fermentas). The ExoI enzyme digests unused amplification primers by degrading

single-stranded DNA. The SAP enzyme inactivates unused dNTP's by removing the 5' phosphate group.

To each PCR reaction were added

0.9 µl of Shrimp Alkaline Phosphatase (1 U/µl)
0.05 µl of Exonuclease I (20 U/µl) and
0.05 µl of deionised ddH_2O

followed by 20 minutes of incubation at 37°C and 15 minutes at 80°C in a Biometra UNO II thermocycler.

Sequencing of the mtDNA Genome

Cycle sequencing reactions were performed using the DYEnamic ET Terminator Cycle Sequencing Kit from Amersham Pharmacia Biotech. In most of the cases internal primers were used for sequencing.

The sequencing mixture with the total volume of 10 µl consisted of:

1 µl	Primer (5pM/µl)
5µl	PCR product
1 µl	DYEnamic™ ET Dye Terminator reagent premix
3µl	2.5x buffer (buffer B 100 µl: 750 mM Tris-HCl pH8.9, 200 mM $(NH_4)_2SO_4$, 01% Tween 20; $MgCl_2$ 30 µl (25mM); ddH_2O 270 µl)

The conditions for sequencing reaction:

Denaturation	94° C	20 s	44-46 cycles
Primer annealing	50-59° C	15 s	
Synthesis	60° C	1 m	

DNA fragments attained during sequencing reaction were precipitated, washed and suspended in 10 µl of MegaBACE loading solution.

Precipitation of Sequencing Products

10 µl or sequencing reaction product was precipitated and purifed according to the following protocol:

2µl of ammonium acetate and dextran added (1.5 M NaCH3COO, pH>8/EDTA
1. (250mM); 20mg/ml red dextran 1:1 mix)
2. 30 µl of 96% ethanol (-20°) added
3. 15-20 minutes at -20°
4. Centrifugation at 13 000 rpm for 15 min
5. After removing the supernatant 200 µl of 70% ethanol (-20°) added for removal of salt ions
6. Centrifugation at 13 000 rpm for 9 min
7. Repeat steps 5. and 6.
8. After removing the supernatant the remaining ethanol is allowed to evaporate
9. The pellet is suspended in 10 µl of MegaBACE loading solution.

Samples were analysed on MegaBACE 1000 capillary sequencer using linear polyacryl amid polymer and 50 cm capillaries.

Data analyses

Sequences were manually aligned with the revised Cambridge Reference Sequence (rCRS) (Andrews *et al.* 1999) in SeqLab (GCG Wisconsin Package 10, Genetics Computer Group). A list of differences compared to the rCRS (Andrews *et al.* 1999) was recorded for each individual. To determin the gene loci MITOMAP web page was used (A Human Mitochondrial Genome Database. http://www.mitomap.org, 2005).
The charcteristics of base changes were assigned with MitoAnalyzer (National Institute of Standards and Technology, Gaithersburg, MD, USA http://www.cstl.nist.gov/biotech/strbase/mitoanalyzer.html, 2000).

Phylogenetic trees were constructed manually according to reduced median and median joining principles (Bandelt *et al.* 1995; Bandelt *et al.* 1999). The time to the most recent common ancestor of each cluster was estimated using ρ, the average transitional distance from the putative root haplotype. For calibration each non-synonymous mutation between

nucleotide positions 3307 15887 was taken equal to 6764 years (Kivisild *et al.* submitted). Standard deviations for time estimates were calculated as in (Saillard *et al.* 2000).

Results and discussion

The nature of mutations detected in the course of the sequencing

All found mutations are listed in supplementary materials tables 2-7. In every haplogroup R sub-cluster many non-synonymous mutations appeared. The highest number of non-synonymous substitutions appears in ATP-synthetase 6 and NADH-dehydrogenase 1 genes. ATP6 gene, which is generally found to be one of the most conserved genes in comparisons between distant species, has the highest amino acid sequence variation of any mitochondrial gene in human mtDNA. (Mishmar *et al.* 2003). In this study there were altogether 8 non-synonymous (ns) substitutions in ATP6 gene. All non synonymous mutations are listed in Table 3.

HG	mutation	type	funcional reagion	AA change	COD pos	AA pos	
R30	**3316**	G->A	ND1	ALA->THR	1	4	
R8	**4205**	T->C	ND1	LEU->SER	2	300	
R2	**4216**	T->C	ND1	TYR->HIS	1	304	**5**
R30	**4225**	A->G	ND1	MET->VAL	1	307	
R30	**4232**	T->C	ND1	ILE->THR	2	309	
R8	**4956**	A->G	ND2	MET->VAL	1	163	
R8	**5062**	C->T	ND2	PRO->LEU	2	198	**4**
R8	**5148**	A->G	ND2	THR->ALA	1	227	
R30	**5442**	T->C	ND2	PHE->LEU	1	325	
R8	**5911**	C->T	COI	ALA->VAL	2	3	
R7	**6136**	T->C	COI	PHE->SER	2	78	**3**
R31	**6480**	G->A	COI	VAL->ILE	1	193	
R8	**7607**	G->A	COII	GLY->SER	1	8	**1**
R6 R30	**8584**	G->A	ATPase6	ALA->THR	1	20	
R5	**8594**	T->C	ATPase6	ILE->THR	2	23	
R31	**8842**	A->G	ATPase6	ILE->VAL	1	106	
general	**8860**	A->G	ATPase6	THR->ALA	1	112	**8**
R6	**8863**	**G->C**	ATPase6	VAL->ILE	1	113	
R5	**8987**	T->C	ATPase6	MET->THR	2	154	
R2 R7	**9110**	T->C	ATPase6	ILE->THR	2	195	
R30	**9142**	T->C	ATPase6	VAL->ILE	1	206	
R5	**10084**	T->C	ND3	ILE->THR	2	9	**1**
R30	**11016**	G->A	ND4L	SER->ASN	2	86	
R8	**11172**	A->G	ND4L	ASN->SER	2	138	**3**
R5	**11409**	C->T	ND4L	PRO->LEU	2	217	
R30	**12406**	G->A	ND5	VAL->ILE	1	24	
R7	**13105**	A->G	ND5	ILE->VAL	1	257	**4**
R31	**13768**	**T->A**	ND5	PHE->ILE	1	478	
R8	**14002**	A->G	ND5	THR->ALA	1	556	
R31	**14553**	C->T	ND6	VAL->ILE	1	41	**1**
R5	**14990**	C->T	Cytb	LEU->PHE	1	82	**2**
general	**15326**	A->G	Cytb	ASN->ASP	1	194	

Table 3. Non synonymous base changes detected ATPase6 – ATP synthase F0 subunit 6; COI – cytochrome c oxidase subunit I; COII – cytochrome c oxidase subunit II; Cytb – Cytochrome b, ND1, 2, 3, 4L, 5, 6 – NADH dehydrogenase subunit 1, 2, 3, 4L, 5, 6;

The coalescence ages of the haplogroups

In coalescence calculations all those mutations listed in the Table 3 were excluded and only synonymous sites were taken into account. Haplogroups differ into several time layers according to their coalescence ages.

R2 37200 ± 8200

R5 34800 ± 10 300

R8 35500 ± 10100

R6 49000 ± 9700

R7 63100 ± 17200

R30 80300 ± 12200

R31 78500 ± 15800

Some age-estimates for haplogroups changed. Coalescence age estimated from coding region mutations for haplogroup R5 gives a more recent time when compared to the estimates from HVS-I (66000 ± 22000). The HVS-I calculation has also much greater standard deviation.

Age estimate for haplogroup R6 based on HVS-I mutations was 30000 ± 11000 ybp, from coding region mutation it was 49000 ± 9700 ybp. This is due to a large amount of non-synonymous coding region mutations which have accumulated in different individuals. In older lineages there are generally less non-synonymous mutations, which often are slightly deleterious and therefore tend to disappear from the gene pool over the longer period of time.

The structure of the phylogenetic tree

All the individual mitochondrial genomes sequenced in the course of this study fitted well into the published phylogeny of R-lineages (Palanichamy *et al.* 2004). However, with the additional information in hand it is possible to define some new sub-clades.

Haplogroup R5

Before full sequences became available, haplogroup R5 was defined by having a mutation in the coding region position 8594 (corresponding to –8592*Mbo*I site) in addition to HVS-I haplotype 16266 16304 (Quintana-Murci *et al.* 2004). Palanichamy *et al* characterized R5 by another three mutations, at 10754, 14544, and 16524. The state of transition 16266

was left open, as it was not sure whether it defines the whole haplogroup or a major sub-clade, since back mutations were observed.

There were three samples from haplogroup R5 sequenced in the course of this study (see Table 4) which all harboured the 16266 transition, this transition is though often seen in other clades and has thus arisen multiple times in haplogroup R, which is the reason one can not use this mutation alone for defining R5.
Among the published sequences there is one individual gone through a back mutation in the position 16266 and additionally lacks the deleted CA nucleotides in the positions 522-523 therefore separating from the rest of the lineages.

State	social status	caste/ religion	language group	Population	sample	HG	HVS1
Sri Lanka	caste	Muslim	Dravidic	Moor	38	R5	266-304-309-325-356
West Bengal	caste	Brahmin	Indo-European	Bengal	46	R5	266-304
Gujarat	caste	N/A	Indo-European	Gujarat	35	R5	266-304-309-325-356

Table 4. Sequenced samples belonging to haplogroup R5

The (16266, 522-523d) clade further branches into two, one major cluster has coding region mutations 11293, 13635, 14040, 15385 in addition to control region mutations 152 and 16356 and comprises the majority of individuals in the tree (10). The other clade comprises three genomes and is separated by transitions 2833, 8987, 9708, 10084 in coding region and 93, 200 motif in the control region. This is a new sub-cluster of the haplogroup. One should though bear in mind, that characterization of more peripheral parts of the tree, which are represented only by a few sequences is not final.

Of the three R5 genomes sequenced in the course of this study two have identical HVS-I regions. It turned out that their coding regions differed from each other by 5 individual mutations. Together they form a separate cluster defined by a common mutation 11409.

The third individual with 16266, 16304 HVS-I sequence differed substantially from the other two by the coding region mutations (the initial separation by HVS-I was thus reliable) (Figure 8).

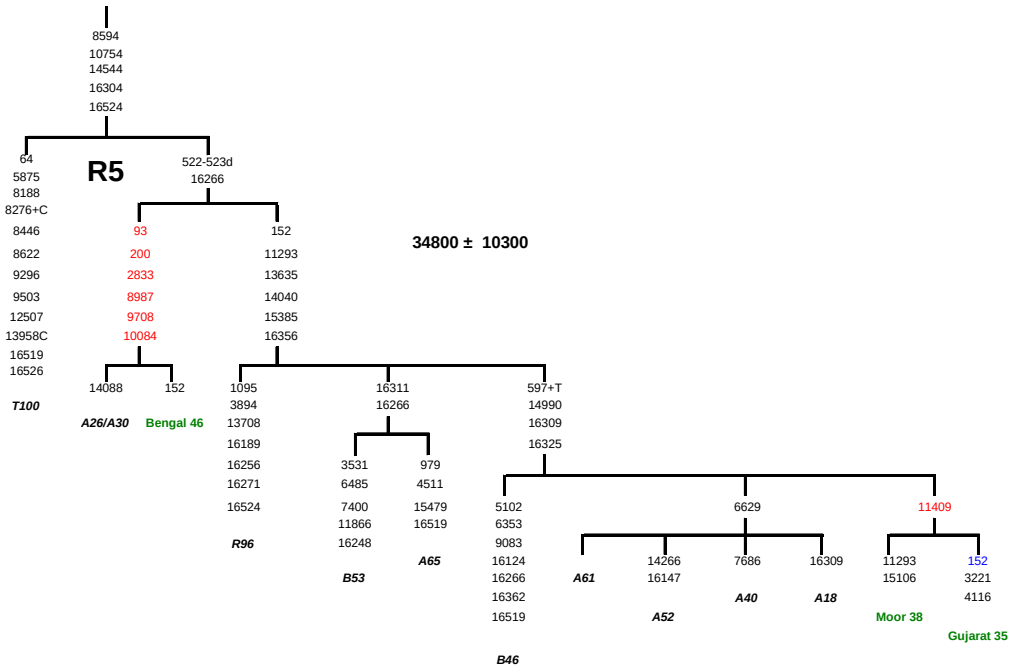


Figure 8. The topology of haplogroup R5. Red colour indicates mutations defining new sub-clades, blue marks back mutations Samples printed in green were sequenced in the course of this study, the rest are from (Palanichamy *et al.* 2004).

Haplogroups R6 and R7.

In the published phylogeny the transition 16362 joins together haplogroups R6 and R7. This mutation is one of the highly variable mutations (Hasegawa *et al.* 1993), it is also seen many times in other haplogroups among R-lineages, therefore it can not serve as a defining mutations for a haplogroup. The coding region mutations of individuals carrying the 16362 transition differ profoundly, separating two haplogroups R6 and R7.

In the coding region the mutation 12285 detected by -12282*Alu*I site as well as HVS-I mutation 16362 (Quintana-Murci *et al.* 2004) remain defining markers for R6. In addition, control region polymorphisms 16274 and 195 appear on the root of the clade. The hypervariable transversion 16129, initially a defining marker, separates in the full sequence tree a sub-cluster (Figure 9).

One individual mitochondrial genome sequenced in the course of this study (Table 5) belonged into haplogroup R6. The sequence of Koya 5 genome formed together with the published sequence T130 a separate cluster, defined by 8584, 11075, 14058 in the coding

region and 228 in control region. All R6 individuals display a high number of individual mutations, corresponding to relatively high age of this clade.

State	social status	language group	Population	sample	HG	HVS1
Andhra Pradesh	tribe	Dravidic	Koya	5	R6	129-266-318-320-362
Andhra Pradesh	tribe	Dravidic	Koya	18	R7	260-261-311-319-362

Table 5. Sequenced samples belonging to R6 and R7 haplogroups.

Another sequenced genome (Koya 18) befitted into haplogroup R7, all defining seven coding region mutations 1442, 6248, 9051, 9110, 10289, 13105, 13830 as well as the control region mutations 16260, 16261, 16319 were confirmed. A new sub-clade of R7 could be defined. It is defined by coding region mutations 1804, 2282, 8557, 12432, 14064 and control region mutations 146 and 16311.

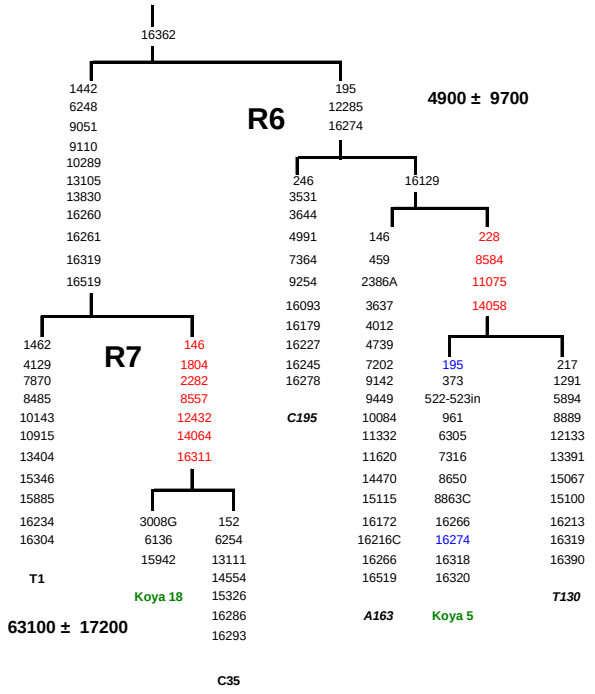


Figure 9. The topology of haplogroups R6 and R7. Colours as in Figure 8.

Haplogroup R8

Haplogroup R8 is defined on the basis of lineages which formed a cluster having common mutations in nucleotide positions 2755, 3384, 7759, 9449, 13215. In the course of this study five lineages were incorporated into the phylogenetic tree (Table 6, Figure 10). Mutations 709, 5911, 9718, 13782 define a sub-cluster of haplogroup R8 encompassing the majority of lineages. Two published sequences in this clade harbour a back mutation 15326, (a CRS specific mutation) which is not present in the four newly sequenced samples, therefore the emergence of this mutation twice was the most parsimonious solution.

State	social status	caste/ religion	language group	Population	sample	HG	HVS1
Maharashtra	caste	Brahmin	Indo-European	Konkanastha Brahmin	41	R8	CRS
Andhra Pradesh	tribe		Dravidic	Koya	30	R8	324
Andhra Pradesh	tribe		Dravidic	Koya	37	R8	324
Andhra Pradesh	tribe		Dravidic	Koya	74	R8	CRS
Andhra Pradesh	tribe		Indo-European	Lambadi	10	R8	292

Table 6. Sequenced samples from haplogroup R8

There is a further division on the basis of transition in the position 8646 which seems to separate another sub-clade, at the moment supported only by two sequences. The other clade is defined by two transitions in the HVS-II region, at nucleotide position 195, 198.

Two samples, Koya 30, Koya 37 (which also have similar HVS-I sequences) formed one cluster separated from other lineages by six coding region and one control region mutation (4205, 7364, 8598, 11172, 11479, 12124 and 16324). Both harbour only one additional individual mutation, exhibiting short time interval from the common ancestor.

Individuals Koya 74 and Cobra 41 also have similar HVS-I sequences, identical with the rCRS, but their coding region separates them by 12 mutations in addition to two HVS-II mutations.

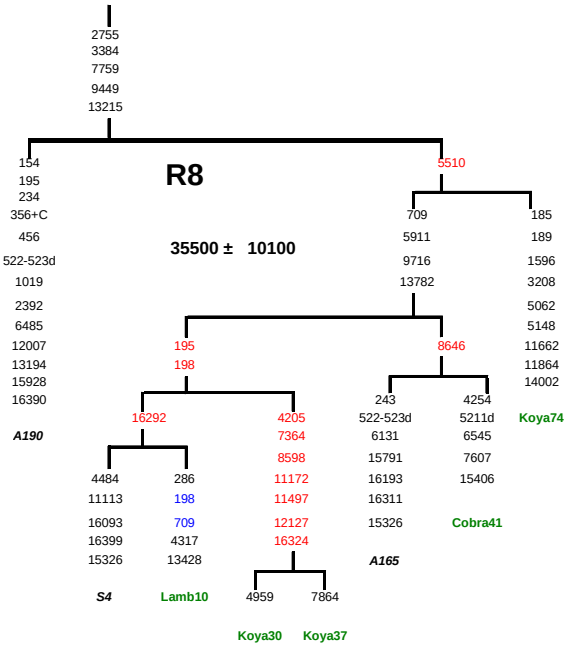


Figure 10. The topology of haplogroup R8. Colours as in Figure 8.

Haplogroup R30

This haplogroup was tentatively defined by the shared mutation at the nucleotide position 8584, which is a recurrent mutation (part of the motive separating a sub-cluster in haplogroup R6). In the published phylogeny there are altogether 5 lineages clustered on the basis of 8584 mutation (Figure 11). There are two separate clusters inside haplogroup R30, which were confirmed by newly sequenced Punjab 47 and Singal 49. The third sample only shares the basal 8584 mutation

State	social status	caste/ religion	language group	Population	sample	HG	HVS1
Sri Lanka	caste	Goyigama	Indo-European	Sinhalese	49	R30	172-278
Punjab	caste	Jat Sikh	Indo-European	Punjab	47	R30	92-189-298-299
Maharashtra	caste	Brahmin	Indo-European	Konkanastha Brahmin	23	R30	286-292

Table 7. Sequenced samples from haplogroup R30

Haplogroup R30 has very long branches indicating its old age. There are two separate sub-clusters both comprising three individuals.

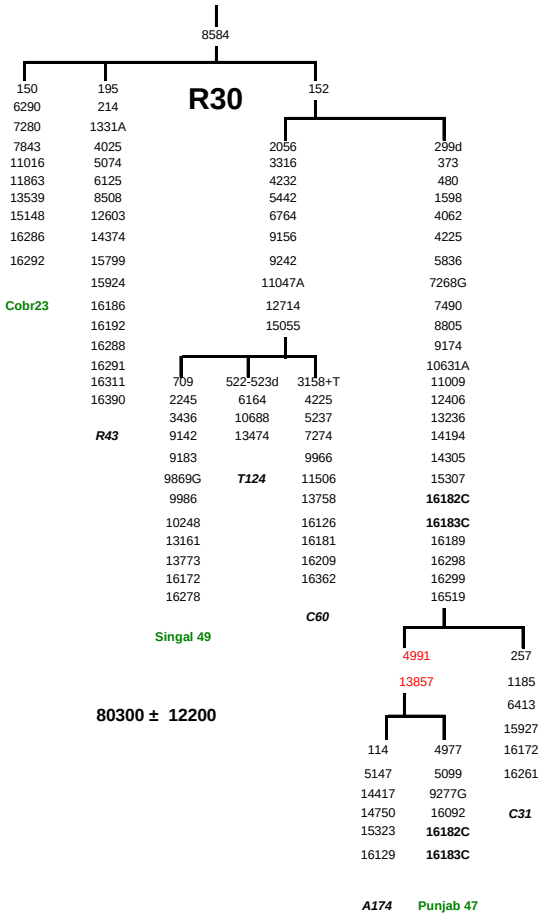


Figure 11. The topology of haplogroup R30. Colours as in Figure 8.

Haplogroup R31

A mutation in the position 15884 clusters lineages into haplogroup R31. Similarly to R30 this haplogroup has two sub-clusters (Figure 12), which display a high number of mutations indicating the deep coalescence age of these sub-clades. Two individual genomes with similar HVS-I sequences (Table 8) were added to the published tree, these sequences are separated from each other by each having one coding region mutation, which indicates a long common history.

State	social status	caste/ religion	language group	Population	sample	HG	HVS1
Rajastan	caste	Kshatriya	Indo-European	Rajput	25	R31	172-304-362
Rajastan	caste	Kshatriya	Indo-European	Rajput	48	R31	172-304-362

Table 8. Sequenced samples from haplogroup R31.

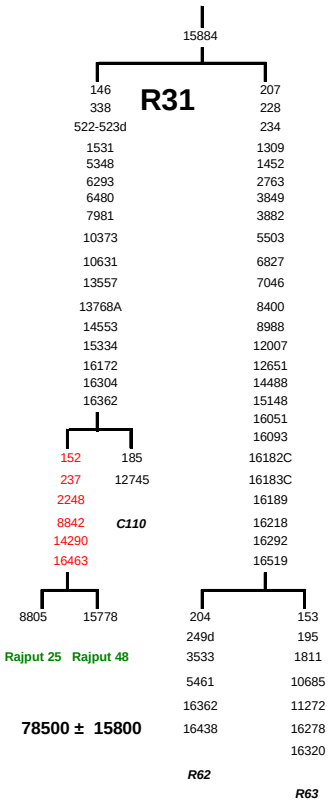


Figure 12. The topology of haplogroup R31. Colours as in Figure 8.

Haplogroup R2

Originally R2 was defined by coding region motif +4216*Nla*III, +4769*Alu*I, -14304*Alu*I and control region mutation 16071 (Quintana-Murci *et al.* 2004). Among published R full sequences (Palanichamy *et al.* 2004) there was only one belonging to haplogroup R2. The mutations 4216, 4769, 14305 and 16071 all were present in that individual.

In this study two additional R2 lineages were sequenced (Table 9). When sequences were incorporated into a tree (Figure 13), R2 cluster is defined by additional nine coding region mutations and five HVS-II mutations. Two genomes sequenced in this study form a separate cluster with shared coding region transitions in nucleotide position 8143, 13434, and a transversion 13914A.

State	social status	caste/ religion	language group	Population	sample	HG	HVS1
Uttar Pradesh	caste	N/A	Indo-European	Uttar Pradesh	606	R2	71
Rajastan	caste	N/A	Indo-European	Rajastan	487	R2	71-293

Table 9. Sequenced samples belonging to haplogroup R2.

When three additional haplogroup R2 lineages from Iran, Georgia and Turkey (unpublished data from Maere Reidla) were incorporated into the phylogenetic tree a clear separation between Indian and non-Indian individuals could be seen. Two coding region mutations separate two clusters – three Indian samples all harbour a transition at nucleotide position 13500, whereas non-Indians have a transition 8027 instead (also 150 and an insertion in the position 303). This could indicate the presence of region-specific clusters in R2. For the affirmation a bigger sample-collection has to be screened for the presence of described mutations.

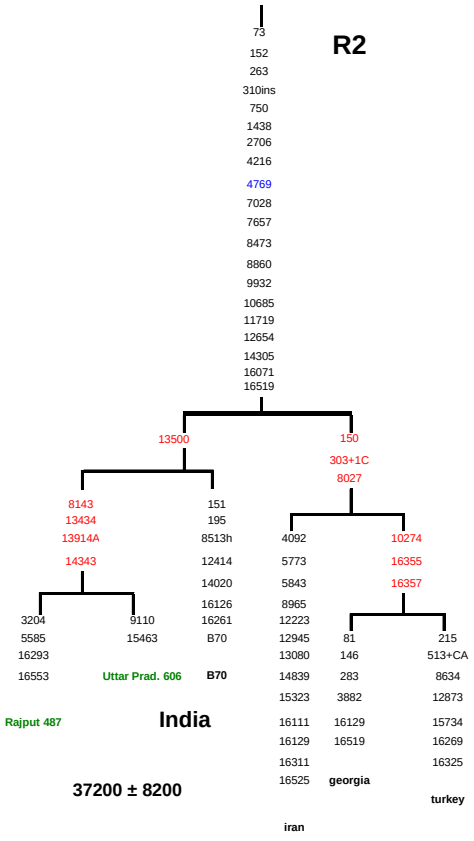


Figure 13. The topology of haplogroup R2 of Indian and non-Indian samples. Colours as in Fig.8.

The fully sequenced mitochondrial genomes belonging to haplogroup R come from individuals from different Indian localities. There are altogether 42 genomes incorporated into the Indian-specific R-lineage topology (for the characteristics of published individuals and the whole tree see Supplementary materials table 8).

Individuals from different regions, social status and linguistic families are spread quite equally between different haplogroups.

14 genomes belong to haplogroup R5, 9 of those are from caste people from Uttar Pradesh in central India. One individual from West-Bengal (eastern India) and one from Gujarat (western India) together with those from Uttar Pradesh all are Indo-European speakers. There are altogether three individuals from southern India (Andhra Pradesh and Sri Lanka) who speak Dravidic languages. One of them is from a tribal Khasi population.

Haplogroup R7 has two Dravidian speakers from Andhra Pradesh, one from a tribal Koya populations and the other from Thagotaveera caste people. The Dravidian speaking tribal forms a clade together with an Indo-European caste individual from northern India.

In the haplogroup R6 two individuals (one tribal one caste) from Andhra Pradesh (Dravidian speaking) cluster together and two Uttar Pradesh caste individuals stay separately.

Haplogroup R8 unites three different language families – four Indo-European speakers, one Austro-Asiatic and three Dravidian speakers. There are caste people from Uttar Pradesh and Maharasthra, tribals from Andhra Pradesh and Meghalaya. One Uttar Pradesh individual is separated from others by 6 mutational steps, just as a tribal from Andhra Pradesh by 5 steps.

8 individuals fit into haplogroup R30. Here three Indo-European speaking caste individuals from central (Uttar Pradesh) and one from northern (Rajastan) India cluster together and have a sister-clade where two individuals from southern India and one from central India are clustered. One caste individual from Maharashtra and one tribal from Andhra Pradesh stem straight from the ancestral node in the tree.

Five individuals have been incorporated into the cluster R31 where two tribals from Andhra Pradesh cluster closely together and two Rajputs form another long-branched cluster, having an Uttar Pradesh caste individual also close by. In this cluster the individuals from southern and northern India are separated by very long branches, indicating a long separate history of these lineages.

It is tempting to draw inferences about geographic or socio-linguistic distribution of the haplogroups based only on this tree, but one has to bear in mind the small sample size represented with this phylogeny and not make too hasty conclusions.

The defined haplogroups can be characterized by exact or close matches in HVS-I motif and diagnostic coding region markers enabling larger scale analyses of the haplogroup frequencies and dispersals. Indian R-samples from the collections of our laboratory and our colleagues from India were screened for such polymorphisms. Previously unassigned R* samples could be classified according to the phylogeny and bigger sample sizes were obtained for the novel haplogroups (unpublished data from Ene Metspalu, Mait Metspalu and Kumarasamy Thangaraj).

Phylogeographic spread of Indian-specific haplogroups.

The frequency distribution of haplogroups R5, R6 and R2 has been described in the literature overview, for those haplogroups the overall picture has not changed. The distribution of haplogroups R7, R8, R30 and R31 is sketched out quite well with the available data. Analyses of the spread of haplogroup R7 has been made with the total sample size of 8216, for other haplogroups the total sample size is 2536.

R7

Haplogroup R7 has a quite intriguing frequency distribution (Figure 14). Not typically for other Indian autochthonous haplogroups it is found significantly more frequently among Austro-Asiatic speaking tribal groups in eastern India. This haplogroup is present also among Indo-European and Dravidic speaking tribals but is absent from the Tibeto Burman speaking tribals. It is rare among the caste people. The high frequency among Austro–Asiatic speaking tribals is coupled to high diversity of haplotypes. The coalescence age for haplogroup R7 is deep (63100 ± 17200). Before making the conclusion that Austro-Asiatics are the oldest inhabitants the diversity in coding region has to be thoroughly investigated between Austro-Asiatics and other populations. Also support from different genetic markers is needed for confirmation because recent findings in archaeology and linguistics suggest the Neolithic spread of Austro-Asiatic languages in connection with rice cultivation in Yangze River basin (Higham 2003).

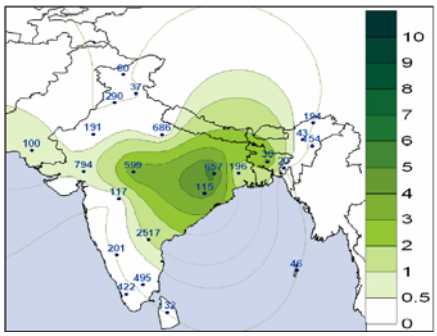


Figure 14. The spread of haplogroup R7 among Indian populations. Haplogroup R7 is more spread among the Austro-Asiatic tribals. Numbers on the map indicate the sample sizes from marked regions.

R8

The highest frequencies of haplogroup R8 (Figure 15) appear in Andra Pradesh together with Sri Lanka and in Rajastan north-western India. The frequency is highest among the Dravidian speaking people, rising to 3% in the far southern India. Haplogroup R8 frequencies among some tribal populations are following - Boksha (4.3 %) from Uttar Pradesh, central India; Koya (9.9 %) and Lambadi (5.8%) who live in Andra Pradesh, south India. Haplogroup R8 has a coalescence age of 35500 ± 10100 which is comparable to the age of M6

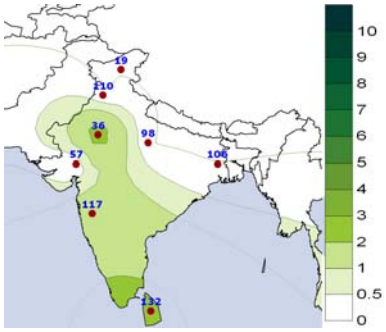


Figure 15. The frequency distribution pattern of haplogroup R8. Only caste people are included in the kriging procedure.

R30

Haplogroup R30 is a very ancient (coalescence time 80300 ± 12200 ybp) haplogroup that prevails in north western, western and central India (Figure 16). The highest frequency is in Gujarat, where the haplogroup is present in caste people. In Punjab this haplogroup is present both in tribal and caste people

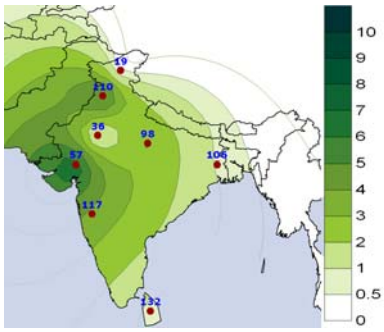
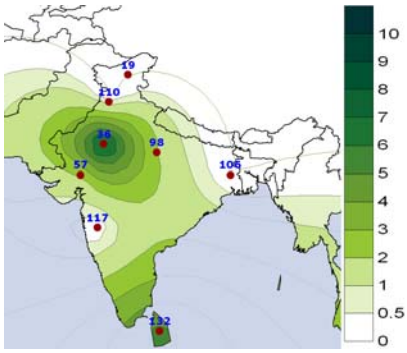


Figure 16. The frequency distribution pattern of haplogroup R30 in India.

Different caste people from Maharstra central India and tribals and caste people from Uttar Pradesh also harbour haplogroup R30. There is also one sample from the northern Himchal Pradesh state and one from Sri Lanka and West Bengal. Among our samples R30 carriers are mainly Indo-European speakers, and there is only one Tibeto-Burman speaker (from Himchal Pradesh). Haplogroup R30 has higher frequencies from north-western to central India, but is present also in other part of the country. The frequency of haplogroup R 30 among tribal populations – Bhoksha 4,3%, Kanet 2,75, Lambadi 1,2%, Lobana 13,9%, Tharu 5,3%.

R31

Haplogroup R31, with coalescence age of 78500 ± 15800, shows higher frequencies in Rajastan on one end of India and Sri Lanka together with the southern tip of the subcontinent on the other. Fewer carriers of this haplogroup inhabit also central India (Uttar Pradesh).



The frequencies in Rajastan and Sri Lanka are quite equal (close to 6%). It is interesting to note that the carriers of this haplogroup are also mostly Indo-European speakers, but also some Dravidic speakers are present. This haplogroup has two different frequency peaks, one in the south and the other in the north. The phylogenetic tree based on fully sequenced genomes from those haplogroups showed distinct separation between individuals from north and south being the only geographic specific clade found in this study.

Conclusions

The main aim of this study was to improve the topology of Indian haplogroup R lineages. The goal was achieved through fully sequencing 17 human mtDNA genomes from Indian individuals. All genomes were integrated into the published phylogeny of haplogroup R (Palanichamy 2004 et al). Additional information was added to the phylogenetic tree, several new sub-clusters emerged.

It was found that the gene with highest amino acid variation is ATP6 as in published data (Mishmar *et al.* 2003), but in the sample set studied here also NADH dehydrogenase 1 gene was found to be with high amino acid variation.

Coalescence ages were calculated for different clusters. Importantly the coalescence ages calculated from coding region mutations changed the time estimates for some haplogroups when compared to previous studies based on HVS-I data.

Haplogroups differed into several time-layers according to their coalescence time. Haplogroups R2, R5 and R8 coalesced ~35000 ypb, haplogroups R6 ~45000 ybp, haplogroup R7 has coalescence time around 55000 ybp. Deepest coalescence time is exhibited by haplogroups R30 and R31 which coalesce ~70000 ybp.

In general, individuals from different geographic localities and socio-cultural backgrounds did not form any distinct cluster on a tree, but there was a deep north-south separation among individual haplotypes in haplogroup R31.

Phylogeographic frequency distribution study was performed with a larger sample set. The frequency distribution of R7, one of the oldest haplogroups among R-derivates, is unlike other autochthonous Indian lineages, more prevailed among Austro-Asiatic speakers.

When the data from Indian genomes belonging to haplogroup R2 was compiled with fully sequenced mtDNA from Near East and Caucasus region genomes a profound difference appeared – individuals from India fell into distinct cluster on the phylogenetic tree. There were two coding region mutations separating the Indian and non-Indian clusters.

Kokkuvõte (Summary in Estonian)

Indias elab tänapäeval üle 1 miljardi inimese, kes moodustavad subkontinendi piires müriaadi erinevaid sotsiaalseid ja lingvistilisi grupeeringuid – Indiat tuntakse kui maailma kõige struktureerituma ühiskonnaga riiki. India geneetilise tausta uurimine annab vastuseid Euraasia asustamise kohta.

Käesolevas töös sekveneeriti täielikult 17 Indiast pärit inimese mitokondriaalse DNA (mtDNA) genoom. Kõik väljavalitud proovid kuulusid superhaplogrupi R* basaalsesse sõlmpunkti ning neid ei olnud algselt võimalik olemasoleva informatsiooni (hüpervariaabel piirkond I ja restriktsiooonifragmetntide pikkuspolümorfismid) põhjal edasi klassifitseerida.

Töö eesmärgiks oli täiendada haplogrupi R fülogeneetilist puud ning anda esmane kirjeldus uute R-alamklastrite levikust Indias.

Töö käigus selgitati välja mitmeid uusi haplogrupi R alamklastreid ning arvutati ka alamhulkade koalestsentsiaegu baseerudes ainult kodeeriva piirkonna sünonüümsetel mutatsioonidel. R-alamhaplogrupid jagunesid koalestsentsiaegade järgi erinevatesse ajaperioodidesse. Haplogrupid R2, R5 ja R8 koalestseeruvad ~35000 aastat tagasi (at), haplogrupp R6 ~45000 at, haplogrupp R7~55000 at. Kõige vanemad haplogrupid on aga R30 ja R31 koalestsentsiajaga 70 000 at.

Erinevatesse puu alamklastritesse jaotunud indiviidid erinesid nii geograafilise kui sotsio-kultuurilise päritolu poolest, vaid haplogrupi R31 siseselt moodustasid Põhja– ja Lõuna-Indiast pärit indiviidid eraldi klastrid.

Haplogruppi R2 kuulunud indiviide võrreldi ka Lähis-Ida ja Kaukaasia päritoluga indiviididega. Fülogeneetilisel puul eralduvad Kesk-Aasia ja India indiviidid eri klastritesse, mistõttu võib oletada lokaalsete alamharude teket selles haplogrupis.

R-alamklastrite sagedusmustritest võib esile tõsta haplogrupi R7 levikut, mis vastupidiselt teistele India-spetsiifilistele haplogruppide esineb kõrgema sagedusega Austro-Aasia keelkonda kuuluvate inimeste seas.

Ka selles töös täielikult sekveneeritud mitokondriaalsed genoomid näitasid, et kõige varieeruvam geen mitokondri DNAs on ATPaas 6 geen, järgnes NADH-dehüdrogenaas 1 geen. Töö tulemuseks on oluliselt täpsustatud R-haplogrupi fülogeneetiline puu, mis on aluseks edasisteks uuringuteks India ja kogu Euraasia mitokondriaalse geenitiigi uurimisele. Töö tulemusena võib öelda ka, et uusi R-alamklastreid tuleb käsitleda võrdsetena teiste, juba hästi läbiuuritud struktuuri ja fülogeograafiaga R-alamhulkadega.

Acknowledgements

I would like to thank my supervisor Ene Metspalu for professional guidance, it has been wonderful to work together with this interesting topic in our field of science. This work would have been impossible to conduct without kind help of Hille Hilpus, Hela Help and Jaan Lind for the technical assistance, specially managing with the sometimes capricious MegaBace 1000. My deepest thanks to Mait Metspalu and Urmas Roostalu for fruitful discussions and a lot of guidance in the writing process of this thesis. I wish to thank all the colleagues and fellow students for the friendly atmosphere and ongoing fruitful discussions.

Last but not least, I thank my family for being always there.

Supplementary materials

Table 1. Primers for PCR and sequencing

Region		Forward		Reverse	
1	**2000**				
		1F.611	CTCCTCAAAGCAATACACTG	1R.1411	TGCTAAATCCACCTTCGACC
		2F.1245	CGATCAACCTCACCACCTCT	23R.5	GAGTGGTTAATAGGGTGATAG
		3F.1854	GGACTAACCCCTATACCTTCTGC	24R.775	AGGCTAAGCGTTTTGAGCTG
		mt520F	CACACACACCGCTGCTAAC	mt836R	TGCTAAAGGTTAATCACTGCTG
2000	**4000**				
		4F.2499	AAATCTTACCCCGCCTGTTT	2R.2007	TGGACAACCAGCTATCACCA
		5F.3169	TACTTCACAAAGCGCCTTCC	3R.2669	GGCAGGTCAATTTCACTGGT
		6F.3796	TGGCTCCTTTAACCTCTCCA	4R.3346	AGGAATGCCATTGCGATTAG
		mt2364F	CTGACAATTAACAGCCCAATATC	5R.3961	ATGAAGAATAGGGCGAAGGG
				mt2484R	GATTTGCCGAGTTCCTTTTACT
4000	**6000**				
		7F.4485	ACTAATTAATCCCCTGGCCC	6R.4654	AAGGATTATGGATGCGGTTG
		8F.5255	CTAACCGGCTTTTTGCCC	7R.5420	CCTGGGGTGGGTTTTGTATG
		mt4155F	CCAACTCATACACCTCCTATG	9F.5855	GAGGCCTAACCCCTGTCTTT
				mt4249R	GAATGCTGGAGATTGTAATGGG
6000	**8000**				
		10F.6469	CTCTTCGTCTGATCCGTCCT	8R.6031	ACCTAGAAGGTTGCCTGGCT
		11F.7148	ACGCCAAAATCCATTTCACT	9R.6642	ATTCCGAAGCCTGGTAGGAT
		12F.7937	ACGAGTACACCGACTACGGC	10R.7315	AGCGAAGGCTTCTCAAATCA
		mt6113F	AATACCCATCATAATCGGAGG	mt6220R	GGTAAGAGTCAGAAGCTTATGT
		mt7925F	GGCGGACTAATCTTCAACTC		
8000	**10000**				
		13F.8621	TTTCCCCCTCTATTGATCCC	11R.8095	CGGGAATTGCATCTGTTTTT
		14F.9230	CCCACCAATCACATGCCTAT	12R.8797	TGGGTGGTTGGTGTAAATGA
		15F.9989	TCTCCATCTATTGATGAGGGTCT	13R.9397	GTGGCCTTGGTATGTGCTTT
		mt9767F	CATTTCCGACGGCATCTA	mt8017R	GAGTACTACTCGATTGTCAACG
10000	**12000**				
		16F.10672	GCCATACTAGTCTTTGCCGC	14R.10130	TGTAGCCGTTGAGTTGTGGT
		17F.11314	TCACTCTCACTGCCCAAGAA	15R.10837	AATTAGGCTGTGGGTGGTTG
		18F.11948	TATCACTCTCCTACTTACAG	16R.11472	TTGAGAATGAGTGTGAGGCG
		mt11614	CATTGCATACTCTTCAATCAGC	mt11748R	GCTAGGCAGAATAGTAATGAGG
12000	**14000**				
		19F.12571	AAACAACCCAGCTCTCCCTAA	17R.12076	GGAGAATGGGGGATAGGTGT
		20F.13338	ACATCTGTACCCACGCCTTC	18R.12772	AGAAGGTTATAATTCCTACG
		mt13539	ATCATACACAAACGCCTGAGC	19R.13507	TCGATGATGTGGTCTTTGGA
				mt13638	TTGACCTGTTAGGGTGAGAAGA
14000	**16000**				
		21F.14000	GCATAATTAAACTTTACTTC	20R.14268	AGAGGGGTCAGGGTTCATTC
		22F.14856	TGAAACTTCGGCTCACTCCT	21R.14998	AGAATATTGAGGCGCCATTG
		23F.15811	TCATTGGACAAGTAGCATCC	22R.15978	AGCTTTGGGTGCTAATGGTG
		24F.16420	CACCATTCTCCGTGAAATCA	mt15431	CGTCTTTGATTGTGTAGTAAGGG
		mt15331F	CCACCTCCTATTCTTGCACG		

Table 2. Mutations detected from haplogroup R2

Raj 487 R2	UP 606 R2	mutation	type	functional region	AA change	COD pos	AA pos
73	73	**73**	A→G	HVS-II			
152	152	**152**	T->C	HVS-II			
263	263	**263**	A->G	HVS-II			
310ins	310ins	**i310**	iC	HVS-II			
750	750	**750**	A->G	12SrRNA			
1438	1438	**1438**	A->G	12SrRNA			
2706	2706	**2706**	A->G	16SrRNA			
3204		**3204**	C->T	16SrRNA			
4216	4216	**4216**	T->C	ND1	TYR->His	1	304
4769A	4769A	**4769**	A->G	ND2		3	
5585		**5585**	G->A	nc			
7028	7028	**7028**	C->T	COI		3	
7657	7657	**7657**	T->C	COII		3	
8143	8143	**8143**	T->C	COII		3	
8473	8473	**8473**	T->C	ATPase8		3	
8860	8860	**8860**	A->G	ATPase6	THR->ALA	1	112
	9110	**9110**	T->C	ATPase6	ILE->THR	2	195
9932	9932	**9932**	G->A	COIII		3	
10685	10685	**10685**	G->A	ND4L		3	
11719	11719	**11719**	G->A	ND4L		3	
12654	12654	**12654**	A->G	ND5		3	
13434	13434	**13434**	A->G	ND5		3	
13500	13500	**13500**	T->C	ND5		3	
13914A	13914A	**13914C->A**		ND5		3	
14305	14305	**14305**	G->A	ND6		3	
14343	14343	**14343**	C->T	ND6	GLY-term	1	111
14766	14766	**14766**	C->T	ND6	ILE->THR	2	7
15326	15326	**15326**	A->G	Cytb	THR->ALA	1	
	15463	**15463**	A->G	Cytb		3	
16071	16071	**16071**	C->T	HVS-I			
16293		**16293**	A->G	HVS-I			
16519	16519	**16519**	T->C	HVS-I			
16553		**16553**	A->G	HVS-I			

Table 3. Mutations detected in haplogroup R5 lineages.

Moor 38 R5	Bengal 46 R5	Gujarat 35 R5	mutation	type	functional region	AA change	COD pos	AA pos
73	73	73	**73**	A→G	HVS-2			
	93		**93**	A->G	HVS-II			
152			**152**	T->C	HVS-II			
	200		**200**	A->G	HVS-II			
263	263	263	**263**	A->G	HVS-II			
ins 315	ins 315	ins315	**i315**	iC	HVS-II			
522-523del	522-523del	522-523del	**522-523d**	del CA	HVS-II			
			522-523i	insCA	HVS-II			
597+T		597+T	**597+T**	ins T	F			
750	750	750	**750**	A->G	12SrRNA			
1438	1438	1438	**1438**	A->G	12SrRNA			
		1961	**1961**	C->T	16SrRNA			
2706	2706	2706	**2706**	A->G	16SrRNA			
	2833		**2833**	C->T	16SrRNA			
		3221	**3221**	A->G	16SrRNA			
		4116	**4116**	C->T	ND1		3	
4769	4769	4769	**4769**	A->G	(ND2)		3	
7028	7028	7028	**7028**	C->T	(COI)		3	
8594	8594		**8594**	T->C	ATPase6	ILE->THR	2	
8860	8860	8860	**8860**	A->G	ATPase6	THR-ALA	1	112
	8987		**8987**	T->C	ATPase6	MET->THR	2	154
	9708		**9708**	T->C	COIII		1	
	10084		**10084**	T->C	ND3	ILE->THR	2	9
10754	10754	10754	**10754**	A->G	ND4L		3	
11293		11293	**11293**	A->G	ND4L		3	
11409		11409	**11409**	C->T	ND4L	PRO->LEU	2	
11719	11719	11719	**11719**	G->A	ND4L		3	
11923			**11923**	A->G	ND4L		3	
13635		13635	**13635**	T->C	ND5		3	
14040		14040	**14040**	G->A	ND5		3	
14544	14544	14544	**14544**	G->A	ND6		1	
14766	14766	14766	**14766**	C->T	ND6	ILE->THR	2	7
14990		14990	**14990**	C->T	Cytb	LEU->PHE	1	82
15106			**15106**	G->A	Cytb		3	
15326	15326	15326	**15326**	A->G	Cytb	THR->ALA	1	
15385		15385	**15385**	C->T	Cytb		3	
16266	16266	16266	**16266**	C->T	HVS-II			
16304	16304	16304	**16304**	T->C	HVS-II			
16309		16309	**16309**	A->G	HVS-II			
16325		16325	**16325**	T->C	HVS-II			
16356		16356	**16356**	T->C	HVS-II			
16519	16519	16519	**16519**	T->C	HVS-II			
16524	16524	16524	**16524**	A->G	HVS-II			

Table 4. Mutations detected in haplogroups R6 and R7.

Ko 18 R7	Ko 5 R6	mutation	type	functional region	AA change	COD pos	AA pos
73	73	**73**	A→G	HVS-II			
146		**146**	T->C	HVS-II			
	228	**228**	G->A	HVS-II			
263	263	**263**	A->G	HVS-II			
	ins 315	**i315**	iC	HVS-II			
	373	**373**	A->G	HVS-II			
		522-523d	del CA	HVS-II			
	522-523in	**522-523i**	insCA	HVS-II			
750	750	**750**	A->G	12SrRNA			
	961	**961**	T->C	12SrRNA			
1438	1438	**1438**	A->G	12SrRNA			
1442		**1442**	G->A	12SrRNA			
1804		**1804**	A->G	16SrRNA			
2282		**2282**	C->T	16SrRNA			
2706	2706	**2706**	A->G	16SrRNA			
3008G		**3008 C->G**		16SrRNA			
	4174C	**4174**	T->C	ND1	TRP->ARG	1	290
4769	4769	**4769**	A->G	ND2		3	
6136		**6136**	T->C	COI	PHE->SER	2	78
6248		**6248**	T->C	COI		3	
	6305	**6305**	G->A	COI		syn3	
7028	7028	**7028**	C->T	COI		syn 3	
	7316	**7316**	G->A	COI		3	
7870C		**7870**	T->C	COII		3	
8557A		**8557**	G->A	ATPase8/ATPase6		3	
	8584	**8584**	G->A	ATPase6	ALA->THR	1	20
	8650	**8650**	C->T	ATPase6		1	
8860	8860	**8860**	A->G	ATPase6	THR->ALA	1	112
	8863C	**8863G->C**		ATPase6	VAL->ILE	1	113
10289		**10289**	A->G	ND3		3	
	11075	**11075**	T->C	NDL4		1	
11719	11719	**11719**	G->A	NDL4		3	
	12285	**12285**	T->C	L			
12432		**12432**	C->T	ND5		3	
13105G		**13105**	A->G	ND5	ILE->VAL	1	
13830		**13830**	T->C	ND5		3	
	14058	**14058**	C->T	ND5		3	
14064		**14064**	C->T	ND5		3	
		14553	C->T	ND6	VAL->ILE	1	41
14766	14766	**14766**	C->T	ND6	ILE-THR	2	7
	15326	**15326**	A->G	Cytb	THR->ALA	1	
15942		**15942**	T->C	T			
	16129	**16129**	G->A	HVS-I			
16260		**16260**	C->T	HVS-I			
16261		**16261**	C->T	HVS-I			
	16266	**16266**	C->T	HVS-I			
16311		**16311**	T->C	HVS-I			
	16318	**16318**	A->G	HVS-I			
16319		**16319**	G->A	HVS-I			
	16320	**16320**	C->T	HVS-I			
16362	16362	**16362**	T->C	HVS-I			
16519	16519	**16519**	T->C	HVS-I			

Table 5. Mutations detected from haplogroup R8.

Lam 10 R8	Ko 30 R8	Ko 37 R8	mutation	type	functional region	AA change	COD pos	AA pos
73	73	73	**73**	A→G	HVS-II			
			185	G->A	HVS-II			
			189	A->G	HVS-II			
195	195	195	**195**	T->C	HVS-II			
	198	198	**198**	C->T	HVS-II			
263	263	263	**263**	A->G	HVS-II			
286			**286**	A->G	HVS-II			
	709	709	**709**	G->A	12SrRNA			
1438	1438	1438	**1438**	A->G	12SrRNA			
			1596	A->G	12SrRNA			
2706	2706	2706	**2706**	A->G	16SrRNA			
2755	2755	2755	**2755**	A->G	16SrRNA			
			3208	C->T	16SrRNA			
3384	3384	3384	**3384**	A->G	ND1		3	
	4205	4205	**4205**	T->C	ND1	LEU->SER	2	300
			4254	T->C	ND1		3	
4317			**4317**	A->G	I			
4769	4769	4769	**4769**	A->G	ND2		3	
	4959		**4956**	A->G	ND2	MET->VAL	1	163
			5062	C->T	ND2	PRO->LEU	2	
			5148	A->G	ND2	THR->ALA	1	227
			5211d (C)		ND2			248
5510	5510	5510	**5510**	A->G	ND2		3	
5911	5911	5911	**5911**	C->T	COI	ALA->VAL	2	3
			6545	C->T	COI		3	
7028	7028	7028	**7028**	C->T	COI		syn 3	
	7364	7364	**7364**	A->G	COI		3	
			7607	G->A	COII	GLY->SER	1	8
7759	7759	7759	**7759**	T->C	COII		3	
	8598	8598	**8598**	T->C	ATPase6		3	
			8646	C->T	ATPase6		3	
9449	9449	9449	**9449**	C->T	COIII		3	
9716	9716	9716	**9716**	T->C	COIII		3	
	11172	11172	**11172**	A->G	ND4L	ASN->SER	2	
	11497	11497	**11497**	C->T	ND4L		3	
			11662	T->C	ND4L		3	
11719	11719	11719	**11719**	G->A	ND4L		3	
			11864	T->C	ND4L		1	
	12127	12127	**12127**	G->A	ND4L		3	
13215	13215	13215	**13215**	T->C	ND5		3	
13428			**13428**	A->G	ND5		3	
13782	13782	13782	**13782**	C->T	ND5		3	
			14002	A->G	ND5	THR->SER	1	556
14766	14766	14766	**14766**	C->T	ND6	ILE->THR	2	7
			15406	C->T	Cytb		3	
16292			**16292**	C->T	HVS-I			
	16324	16324	**16324**	T->C	HVS-I			
16519	16519	16519	**16519**	T->C	HVS-I			

Table 6. Mutations detected in haplogroup R30 (to be continued on the next page)

Pu 47 R30	Sin 49 R30	Cob 23 R30	mutation	type	functional region	AA change	COD pos	AA pos
73	73	73	**73**	A→G	HVS-II			
		150	**150**	C->T	HVS-II			
152	152		**152**	T->C	HVS-II			
263	263	263	**263**	A->G	HVS-II			
299 del			**299d**	dC	HVS-II			
ins 309			**i309**	iC	HVS-II			
ins 315			**i315**	iC	HVS-II			
373			**373**	A->G	HVS-II			
480			**480**	T->C	HVS-II			
	709		**709**	G->A	12SrRNA			
750			**750**	A->G	12SrRNA			
1438	1438	1438	**1438**	A->G	12SrRNA			
1598			**1598**	G->A	12SrRNA			
	2056		**2056**	G->A	16SrRNA			
	2245		**2245**	A->G	16SrRNA			
2706	2706	2706	**2706**	A->G	16SrRNA			
	3316		**3316**	G->A	ND1	ALA->THR	1	4
	3436		**3436**	G->A	ND1	GLY->THR	1	
4062			**4062**	T->C	ND1		3	
4225			**4225**	A->G	ND1	MET->VAL	1	307
	4232		**4232**	T->C	ND1	ILE->THR	2	309
4769	4769	4769	**4769**	A->G	ND2		3	
4977			**4977**	T->C	ND2		1	
4991			**4991**	G->A	ND2		3	
5099			**5099**	C->T	ND2		3	
	5442		**5442**	T->C	ND2	PHE->LEU	1	325
5836			**5836**	A->G	Y			
		6290	**6248**	T->C	COI		3	
	6764		**6764**	G->A	COI		3	
7028	7028	7028	**7028**	C->T	COI		3	
7268G			**7268 T->G**		COI		3	
		7280	**7280**	C->T	COI		3	
7490			**7490**	A->G	S			
	7864		**7864**	C->T	COII		3	
		7843	**7843**	A->G	COII		3	
8584	8584	8584	**8584**	G->A	ATPase6	ALA->THR	1	20
8805			**8805**	A->G	ATPase6		3	
8860			**8860**	A->G	ATPase6	THR->ALA	1	112
	9142		**9142**	T->C	ATPase6	VAL->ILE	1	206
	9156		**9156**	A->G	ATPase6		3	
9174			**9174**	T->C	ATPase6		3	
	9183		**9183**	C->T	ATPase6		3	

Table 6. (continued) Mutations detected from the haplogroup R30

Pu 47 R30	Sin 49 R30	Cob 23 R30	mutation	type	functional region	AA change	COD pos	AA pos
	9242		**9242**	A->G	COCIII		3	
9277G			**9277C->G**		COCIII		3	
	9869G		**9869C->G**		COCIII		3	
	9986		**9986**	G->A	COCIII		3	
	10248		**10248**	T->C	ND3		1	
10631A			**10631C->A**		ND4L		3	
11009			**11009**	T->C	ND4L		1	
		11016	**11016**	G->A	ND4L	SER->ASN	2	
	11047A		**11047C->A**		ND4L		3	
11719	11719	11719	**11719**	G->A	ND4L		3	
		11863	**11863**	C->T	ND4L		3	
12406			**12406**	G->A	ND5	VAL->ILE	1	24
	12714		**12714**	T->C	ND5		3	
	13161		**13161**	T->C	ND5		3	
13236			**13236**	A->G	ND5		3	
		13539	**13539**	A->G	ND5		3	
	13773		**13773**	A->G	ND5		3	
13857			**13857**	A->G	ND5		3	
14194			**14194**	C->T	ND6		3	
14305			**14305**	G->A	ND6		3	
14766	14766	14766	**14766**	C->T	ND6	ILE->THR	2	7
	15055		**15055**	T->C	Cytb		3	
		15148	**15148**	G->A	Cytb		3	
15307			**15307**	C->T	Cytb		3	
15326	15326	15326	**15326**	A->G	Cytb	THR->ALA	1	
16092			**16092**	T->C	HVS-I			
	16172		**16172**	T->C	HVS-I			
16189			**16189**	T->C	HVS-I			
	16278		**16278**	C->T	HVS-I			
		16286	**16286**	C->T	HVS-I			
		16292	**16292**	C->T	HVS-I			
16298			**16298**	T->C	HVS-I			
16299			**16299**	A->G	HVS-I			
	16519	16519	**16519**	T->C	HVS-I			

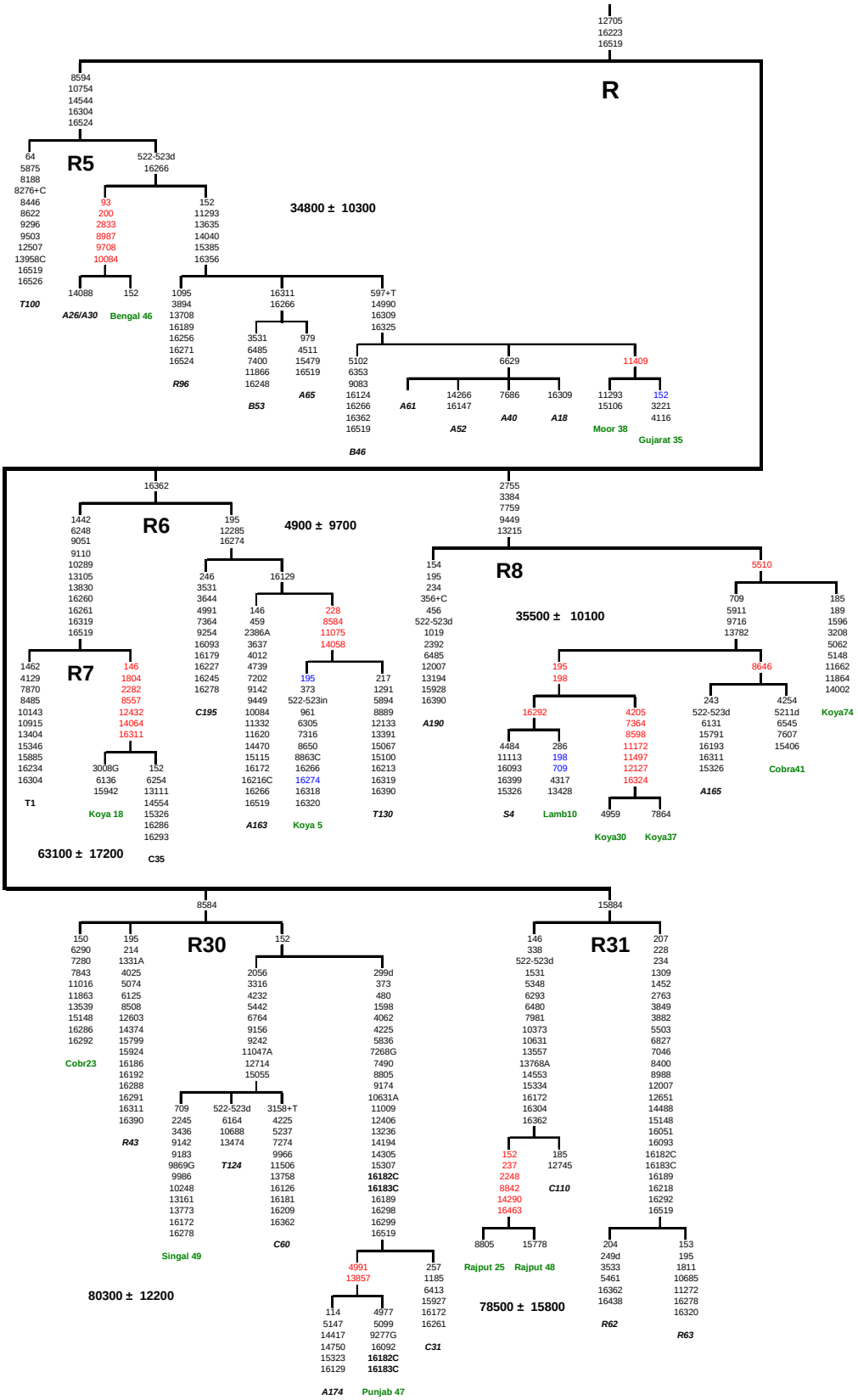
Table 7. Mutations detected in haplogroup R31.

Raj 25 R31	Raj 48 R31	mutation	type	functional region	AA change	COD pos	AA pos
73	73	73	A→G	HVS-II			
146	146	146	T->C	HVS-II			
152	152	152	T->C	HVS-II			
237	237	237	A->G	HVS-II			
263	263	263	A->G	HVS-II			
338	338	338	C->T	HVS-II			
522-523d	522-523d	522-523d	del CA	HVS-II			
1438	1438	1438	A->G	12SrRNA			
1531	1531	1531	C->T	12SrRNA			
2248	2248	2248	T->C	16SrRNA			
2706	2706	2706	A->G	16SrRNA			
4769	4769	4769	A->G	ND2		3	
5348	5348	5348	C->T	ND2		3	
6293	6293	6293	T->C	COI		3	
6480	6480	6480	G->A	COI	VAL->ILE	1	193
7028	7028	7028	C->T	COI		3	
7981	7981	7981	C->T	COII		3	
8805		8805	A->G	ATPase6		3	
8842		8842	A->G	ATPase6	ILE->VAL	1	106
10373	10373	10373	G->A	ND3		3	
10631	10631	10631	C->T	ND4L		3	
11719	11719	11719	G->A	ND4L		3	
13557	13557	13557	A->G	ND5		3	
13768A	13768A	13768T->A		ND5	PHE->ILE	1	478
14290	14290	14290	T->C	ND6		3	
14553	14553	14553	C->T	ND6	VAL->ILE	1	41
14766	14766	14766	C->T	ND6	ILE->THR	2	7
15326	15326	15326	A->G	Cytb	THR->ALA	1	
15334	15334	15334	C->T	Cytb		3	
	15778	15778	C->T	Cytb		3	
15884	15884	15884	G->A	HVS-I			
16172	16172	16172	T->C	HVS-I			
16304	16304	16304	T->C	HVS-I			
16362	16362	16362	T->C	HVS-I			
16463	16463	16463	A->G	HVS-I			
16519	16519	16519	T->C	HVS-I			

Table 8. Characteristics of published R-haplogroup samples.

HG	sample ID	state	cast/tribal	language group	population	HVS-I (-16000)
R5	T100	Andra Pradesh	cast	Dravidian	Thogataveera	304 524 526
R5	A26/A30	Uttar Pradesh	cast	Indo-European	Bhargava	266 304 524
R5	R96	Andra Pradesh	tribal	Dravidian	Reddy	189 256 266 271 304 356
R5	B53	Uttar Pradesh	cast	Indo-European	Chaturvedi	248 304 311 356 524
R5	B65	Uttar Pradesh	cast	Indo-European	Chaturvedi	304 311 356 524
R5	B46	Uttar Pradesh	cast	Indo-European	Chaturvedi	124 304 309 325 356 362
R5	A61	Uttar Pradesh	cast	Indo-European	Bhargava	266 304 309 325 356
R5	A52	Uttar Pradesh	cast	Indo-European	Chaturvedi	147 304 309 325 356
R5	A40	Uttar Pradesh	cast	Indo-European	Chaturvedi	266 304 309 325 356
R5	A18	Uttar Pradesh	cast	Indo-European	Chaturvedi	266 304 325 356
R6	C195	Uttar Pradesh	cast	Indo-European	Brahmin	93 179 227 245 274 278 362
R6	A163	Uttar Pradesh	cast	Indo-European	Bhargava	129 172 216C 266 274 362
R6	T130	Andra Pradesh	cast	Dravidian	Thogataveera	129 213 274 319 362 390
R7	T1	Andra Pradesh	cast	Dravidian	Thogataveera	234 260 261 304 319 362
R7	C35	Uttar Pradesh	cast	Indo-European	Brahmin	260 261 286 293 311 319 362
R8	A190	Uttar Pradesh	cast	Indo-European	Bhargava	390
R8	S4	Meghalaya	tribal	Austro-Asiatic	Khasi	93 292 399
R8	A165	Uttar Pradesh	cast	Indo-European	Bhargava	193 311
R30	R43	Andra Pradesh	tribal	Dravidian	Reddy	186 192 288 291 311 390
R30	T124	Andra Pradesh	cast	Dravidian	Thogataveera	172 278
R30	C60	Uttar Pradesh	cast	Indo-European	Chaturvedi	126 181 209 362
R30	A174	Uttar Pradesh	cast	Indo-European	Bhargava	129 182C 183C 189 298 299
R30	C31	Uttar Pradesh	cast	Indo-European	Brahmin	172 182C 183C 189 261 298 299
R31	C110	Uttar Pradesh	cast	Indo-European	Brahmin	172 304 362
R31	R62	Andra Pradesh	tribal	Dravidian	Reddy	51 93 182C 183C 189 218 292 362 438
R31	R63	Andra Pradesh	tribal	Dravidian	Reddy	51 93 182C 183C 189 218 278 292 320

Figure 1. Autochthonous South Asian R lineages



References

Alves-Silva J, da Silva Santos M, Guimarães PE, Ferreira AC, Bandelt H-J, Pena SD, Prado VF (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67:444-461

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 290:457-465

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

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

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, 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

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BV, Rasanayagam A, Hammer MF (1998) Female gene flow stratifies Hindu castes. Nature 395:651-652

Bandelt HJ, Alves-Silva J, Guimaraes PE, Santos MS, Brehm A, Pereira L, Coppa A, Larruga JM, Rengo C, Scozzari R, Torroni A, Prata MJ, Amorim A, Prado VF, Pena SD (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. Ann Hum Genet 65:549-563.

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

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

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya N, Roychoudhury S, Majumder P (2003) Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res 13:2277-2290

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31-36

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

Chen X, Prosser R, Simonetti S, Sadlock J, Jagiello G, Schon EA (1995a) Rearranged mitochondrial genomes are present in human oocytes. Am J Hum Genet 57:239-247.

Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC (2000) mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am J Hum Genet 66:1362-1383

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995b) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133-149

Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, Yang ZQ, Lin KQ, Li P, Wu M, Geng ZC, Tan CC, Du RF, Jin L (1998) Genetic relationship of populations in China. Proc Natl Acad Sci U S A 95:11763-11768

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

Cordaux R, Saha N, Bentley G, Aunger R, Sirajuddin S, Stoneking M (2003) Mitochondrial DNA analysis reveals diverse histories of tribal populations from India. Eur J Hum Genet 3:253-264

Derbeneva OA, Starikovskaia EB, Volod'ko NV, Wallace DC, Sukernik RI (2002a) [Mitochondrial DNA variation in Kets and Nganasans and the early peoples of Northern Eurasia]. Genetika 38:1554-1560.

Derbeneva OA, Sukernik RI, Volodko NV, Hosseini SH, Lott MT, Wallace DC (2002b) Analysis of mitochondrial DNA diversity in the Aleuts of the Commander Islands and its implications for the genetic history of Beringia. Am J Hum Genet 71:415-421

Diamond JM (1988) Express train to Polynesia. Nature 336:307-308

Endicott P, Gilbert M, Stringer C, Lalueza-Fox C, Willerslev E, Hansen A, Cooper A (2003) The genetic origins of the Andaman Islanders. Am J Hum Genet 72:178-184

Finnilä S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68:1475-1484.

Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935-945

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

Gadgil M, Joshi, N.V., Shambu Prasad,U.V., Manoharan,S., Suresh, Patil (1997) Peopling of India. In: Rao BaNA (ed) The Indian Human Heritage. Universities Press, Hyderabad, India, pp pp.100-129

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A 77:6715-6719

Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, Excoffier L (1995) Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Mol Biol Evol 12:334-345

Grun R, Huang PH, Huang W, McDermott F, Thorne A, Stringer CB, Yan G (1998) ESR and U-series analyses of teeth from the palaeoanthropological site of Hexian, Anhui Province, China. J Hum Evol 34:555-564

Hasegawa M, Di Rienzo A, Kocher TD, Wilson AC (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. J Mol Evol 37:347-354

Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70:1152-1171.

Higham C (2003) Languages and Farming Dispersals: Austroasiatic Languages and Rice Cultivation. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. The McDonald Institute for Archaeological Research, Cambridge

Horai S, Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S, Park KS, Omoto K, Pan IH (1996) mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. Am J Hum Genet 59:579-590

Ingman M, Gyllensten U (2001) Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J Hered 92:454-461.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408:708-713

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 T, Kaldma K, Metspalu M, Parik J, Papiha SS, Villems R (1999b) The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World. In: Papiha SS, Deka R, Chakraborty R (eds) Genomic diversity. Kluwer Academic/Plenum Publishers, pp 135-152

Kivisild T, Papiha SS, Rootsi S, Parik J, Kaldma K, Reidla M, Laos S, Metspalu M, Pielberg G, Adojaan M, Metspalu E, Mastana SS, Wang Y, Gölge M, Demirtas H, Schnekenberg E, Stefano GF, Geberhiwot T, Claustres M, Villems R (2000) An Indian ancestry: a key for understanding human diversity in Europe and beyond. In: Renfrew C, Boyle K (eds) Archaeogenetics: DNA and the population prehistory of Europe. McDonald Institute for Archaeological Research University of Cambridge, Cambridge, pp 267-279

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752-770

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk H-V, Stepanov V, Gölge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R (2003a) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72:313-332

Kivisild T, Rootsi S, Metspalu M, Metspalu E, Parik J, Kaldma K, Usanga E, Mastana S, Papiha S, Villems R (2003b) The genetics of language and farming spread in India. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. The McDonald Institute for Archaeological Research, Cambridge, pp 215-222

Kivisild T, Tolk H-V, Parik J, Wang Y, Papiha SS, Bandelt H-J, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 19:1737-1751 (erratum 1720:1162)

Kolman CJ, Bermingham E (1997) Mitochondrial and nuclear DNA diversity in the Choco and Chibcha Amerinds of Panama. Genetics 147:1289-1302

Kong Q-P, Yao Y-G, Sun C, Bandelt H-J, Zhu C-L, Zhang Y-P (2003) Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 73:671-676

Lahr M, Foley R (1994) Multiple dispersals and modern human origins. Evol Anthropol 3:48-60

Lightowlers RN, Chinnery PF, Turnbull DM, Howell N (1997) Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. Trends Genet 13:450-455.

Loogväli EL, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, Metspalu E, Tambets K, et al. (2004) Disuniting Uniformity: A Pied Cladistic Canvas of mtDNA Haplogroup H in Eurasia. Mol Biol Evol

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

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034-1036

Macaulay VA, Richards MB, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonné-Tamir B, Sykes B, Torroni A (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232-249

Majumder PP (2001a) Ethnic populations of India as seen from an evolutionary perspective. J Biosci 26:533-545.

Majumder PP (2001b) Indian caste origins: genomic insights and future outlook. Genome Res 11:931-932.

Malyarchuk BA, Rogozin IB, Berikov VB, Derenko MV (2002) Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region. Hum Genet 111:46-53

McDougall I, Brown FH, Fleagle JG (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433:733-736

Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MTP, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R (2004) Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet 5:26

Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci USA 100:171-176

Nei M, Roychoudhury A (1993) Evolutionary relationships of human populations on a global scale. Mol Biol Evol 10:927-943

Palanichamy M, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, Khan F, Wang C-Y, Chaudhuri T, Palla V, Zhang Y-P (2004) Phylogeny of mtDNA macrohaplogroup N in India based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75:966-978

Papiha SS (1996) Genetic variation in India. Hum Biol 68:607-628

Quintana-Murci L, Chaix R, Wells S, Behar D, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti A, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Mehdi Q, Torroni A, McElreavey K (2004) Where West meets East: The complex mtDNA landscape of the Southwest and Central Asian corridor. Am J Hum Genet 74:827-845

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

Rajkumar R, Banerjee J, Gunturi HB, Trivedi R, Kashyap VK (2005) Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evol Biol 5:26

Reidla M, Kivisild T, Metspalu E, Kaldma K, Tambets K, Tolk H, Parik J, et al. (2003) Origin and Diffusion of mtDNA Haplogroup X. Am J Hum Genet 73:1178-1190

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251-1276

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.

Roychoudhury S, Roy S, Dey B, Chakraborty M, Roy M, Roy B, Ramesh A, Prabhakaran N, Rani MVU, Vishwanathan HM, M., Sil SKM, P. P. (2000) Fundamental genomic unity of ethnic India is revealed by analysis of mitochondrial DNA. Current Science 79:1182-1192

Saillard J, Evseva I, Tranebjaerg L, Norby S (2000) Mitochondrial DNA diversity among Nenets. In: Renfrew C, Boyle K (eds) Archaeogenetics: DNA and and the population prehistory of Europe. McDonald Institute for Archaeological Research Monograph Series, Cambridge University, Cambridge, pp 255-258

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. Am J Hum Genet 71:1082-1111.

Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74:454-465. Epub 2004 Feb 2010.

Sauer C (1962) Seashore - primitive home of man? Proc Am Phil Soc 106:41-47

Schurr TG, Wallace DC (2002) Mitochondrial DNA diversity in Southeast Asian populations. Hum Biol 74:431-452.

Singh KS (ed) (1997) The Scheduled Tribes. Vol III. Oxford University Press, Oxford

Starikovskaya YB, Sukernik RI, Schurr TG, Kogelnik AM, Wallace DC (1998) mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of Ancient Beringia and the peopling of the New World. Am J Hum Genet 63:1473-1491

Stoneking M (2000) Hypervariable sites in the mtDNA control region are mutational hotspots. Am J Hum Genet 67:1029-1032.

Stoneking M, Jorde LB, Bhatia K, Wilson AC (1990) Geographic variation in human mitochondrial DNA from Papua New Guinea. Genetics 124:717-733

Stringer C (2000) Coasting out of Africa. Nature 405:24-25, 27

Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza L, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last ice age. Am J Hum Genet 65:1718-1724

Sutovsky P, Van Leyen K, McCauley T, Day BN, Sutovsky M (2004) Degradation of paternal mitochondria after fertilization: implications for heteroplasmy, assisted reproductive technologies and mtDNA inheritance. Reprod Biomed Online 8:24-33.

Taanman JW (1999) The mitochondrial genome: structure, transcription, translation and replication. Biochim Biophys Acta 1410:103-123.

Tambets K, Rootsi S, Kivisild T, Help H, Serk P, Loogvali EL, Tolk HV, et al. (2004) The Western and Eastern Roots of the Saami--the Story of Genetic "Outliers" Told by Mitochondrial DNA and Y Chromosomes. Am J Hum Genet 74:661-682

Tanaka M, Cabrera VM, Gonzalez AM, Larruga JM, Takeyasu T, Fuku N, Guo L-J, et al. (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14:1832–1850

Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA, Singh L (2005) Reconstructing the origin of Andaman Islanders. Science 308:996

Torroni A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, Forster P, Savontaus ML, Bonne-Tamir B, Scozzari R (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62:1137-1152

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, Martinez-Cabrera V, et al. (2001a) A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet 69:844-852.

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

Torroni A, Neel JV, Barrantes R, Schurr TG, Wallace DC (1994) Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc Natl Acad Sci USA 91:1158-1162

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001b) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69:1348-1356.

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

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613-623

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238:211-230

Wallace DC, Shoffner JM, Trounce I, Brown MD, Ballinger SW, Corral-Debrinski M, Horton T, Jun AS, Lott MT (1995) Mitochondrial DNA mutations in human degenerative diseases and aging. Biochim Biophys Acta 1271:141-151

Walter RC, Buffler RT, Bruggemann JH, Guillaume MM, Berhe SM, Negassi B, Libsekal Y, Cheng H, Edwards RL, von Cosel R, Neraudeau D, Gagnon M (2000) Early human occupation of the Red Sea coast of Eritrea during the last interglacial. Nature 405:65-69.

Watson E, Forster P, Richards M, Bandelt HJ (1997) Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 61:691-704

Yao Y-G, Kong Q-P, Bandelt H-J, Kivisild T, Zhang Y-P (2002) Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70:635-651

Yao Y-G, Zhang Y-P (2002) Phylogeographic analysis of mtDNA variation in four ethnic populations from Yunnan Province: new data and a reappraisal. J Hum Genet 47:311-318