CHAPTER ONE

Introduction:

South Asia, constituting mainly the Indian sub-continent, is a panorama of social diversities, racial differences and rich cultural heritage. However many facets of history, race and culture of this area are still relatively unexplored (Bhasin , 2006).

Genomic reconstruction of migration trails of modern humans indicates that south Asia has served as a major corridor for the geographic dispersal of humans from out-of- Africa. The extant population of south Asia comprises a large number of isolated groups of varying sizes. This, coupled with the facts that Indian populations have had early contacts with populations outside of the region and that there have been many invasions of India, has resulted in a high degree of genetic diversity in this region (Majumder, 2008).

The populations of India and other South Asian countries offer great opportunities to study socio-cultural and genetic variability. Perhaps, nowhere in the world people in a small geographic area are distributed as such a large number of ethnic, castes, religious and linguistic groups as in India and other South Asian countries. All these groups are not entirely independent; people belong concurrently to two or more of these groups. People of different groups living side by side for hundreds or even thousands of years try to retain their separate entities by practicing endogamy.

The intriguing aspect of the Indian populations is the immense cultural and biological diversity of the various castes, tribes and other ethnic groups (e.g. religious communities). Comprehending the possible reasons for this extensive diversity and tracing the history of the populations, viz., the origin, migration and settlement, has been the focus of study for archeologists, historians, linguists, anthropologists and of late for molecular geneticists.

Archeological and genetic studies have revealed that India occupies an important geographical location and has served as a major corridor for the dispersal of modern humans (Cavalli-Sforza *et al,* 1994; Majumder, 1998; Cann, 2001). From the available prehistoric, ethno-historic, linguistic and biological evidences, it appears that the Indian subcontinent was probably first peopled approximately 60,000 years BP (Misra, 1992; Kivisild *et al,* 2003). Though the entire Indian subcontinent shows diverse ethnic composition, Eastern and North eastern part of India covering the sub-Himalayan terrain has been a hotspot for linguistic (Ruhlen, 1991) and population diversity and thought to have witnessed and retained the signature many ancient and contemporary

genetic event of population migration and settlement. The north-eastern and sub-Himalayan region of India is predominantly occupied by speakers of the Tibeto-Burman languages (of the Sino-Tibetan family) along with Austro-Asiatic (AA) and Indo European (IE) speakers along with sporadic presence of Dravidian group. There are two contrasting and competing models based on the mtDNA and Y-chromosome studies in the NE populations regarding the possible role of NE India in the population migration as well as in settlement. One group based on their mitochondrial DNA (mtDNA) and Y-chromosome studies inferred that Northeast India might have acted as a barrier rather than the facilitator of the movement of populations both into and out of India (Courdax *et al,* 2004) where as the same view has summarily been rejected in a recent study by Reddy *et al*, 2007 who confirms that Northeast India has acted as a corridor for initial movement of populations.

It is now well defined that people belonging to the Indo-European, Dravidian, and Tibeto-Burman groups represent both caste and tribes, but the Austro-Asiatic speakers are exclusively tribal. Though the tribal populations of India possess higher genetic diversity than caste populations and are considered as the possible original inhabitants of India (Thapar 1966; Basu *et al*, 2003, The Indian Genome variation Consortium, 2008), but their evolutionary histories with respect to settlement, biological origin and relationships with caste populations have been debated comprehensivel and remained as an enigma till to date. The phylogenetic relationship between tribe and castes have been shown quite explicitly in the Southern part of India (Kivisild *et al,* 2003, Cordaux *et al*, 2004) however, the tribe-caste relationship with respect to uniparental markers have not yet been studied in Eastern and North Eastern India which is the site where TB and AA speakers are abundant and thought to have made an important contribution in the sub-Himalayan gene pool. Therefore, it is crucial to carry out such genetic investigations in culturally and linguistically disparate and ethnically well-defined populations inhabiting the Eastern and North Eastern periphery of sub-Himalayan region.

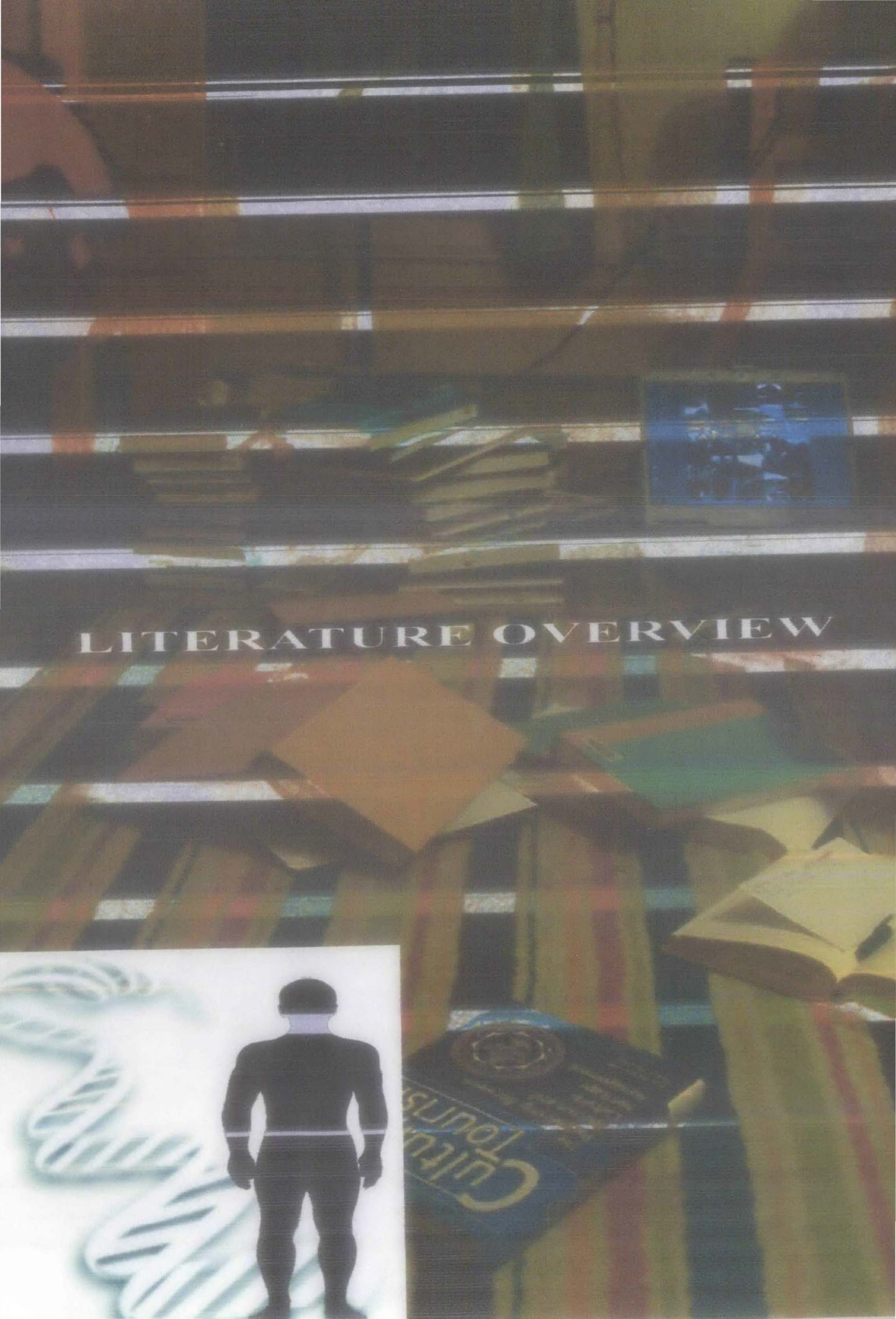
Reconstruction of the process of peopling is problematic using unlinked autosomal (biparental) markers. Mitochondrial DNA (mtDNA) and Y-chromosomal markers have proved to be very useful in reconstructing patterns and tracing trails of human migration, in spite of the limitation that these genomic regions do not undergo recombination and hence

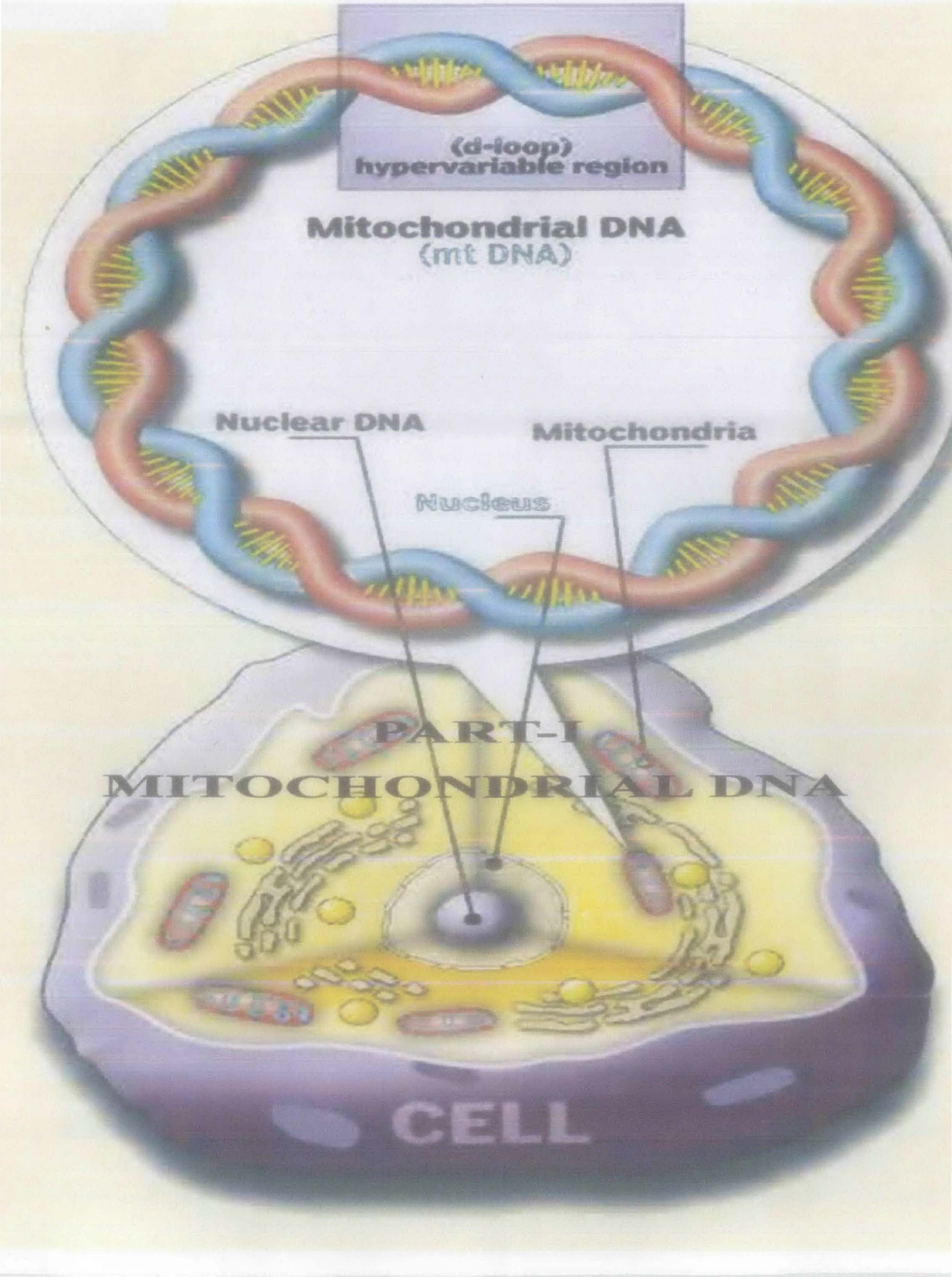
are essentially transmitted as single markers. Another limitation of inferences based on mtDNA and Y-chromosomal studies is that since these are uniparental markers, the impact of genetic drift on these marker systems is greater because of reduced effective population size compared to biparental (autosomal) markers. In spite of these limitations, uniparental markers have provided considerable information for tracing trails of human migration (Majumder, 2008).

However, a far better and vivid picture is expected to emerge by a stratified sampling strategy so as to represent large set of well-defined ethnic groups from wider geographical regions. Increased molecular resolution by employment of large set of markers and application of appropriate statistical tools for analysis of the results would assist in resolving the persisting enigma regarding populations of Indian sub-continent. In addition, the population based molecular genetic studies will also enhance our understanding of the health and disease aspect of tribes and castes in India, leading to the effective control of region and population specific diseases.

The present investigation has been carried out in sub-Himalayan populations from East/North East India, Nepal and Bhutan belonging to different linguistic affiliations, ethnicity and religious group to further understand i) the role of North East Indian passage way in population migration, ii) to delineate to relationships between caste and tribes in sub-Himalayan and iii) to address issues of population stratification and genetic susceptibility to diseases and the influence of geographic, linguistic and religious factors on population diversity in this region , iv) to understand the genomic diversity amongst various groups in North east India and their relation with other Indian as well as Asian populations by employing high resolution phylogeny and phylogeography analysis of mitochondrial genome and SNP markers of NRY region of Y-chromosome.

LITERATURE OVERVIEW





PART-I

MITOCHONDRIAL DNA

Literature overview:

Part-I:

1. Mitochondria

The mitochondrion is an organelle located in the cytoplasm of the cell and is thought to have originated from endosymbiosis of a proteobacteria into a primitive eukaryotic cell. Mitochondria are the cell's power producers. They convert energy into forms that are usable by the cell. They are the sites of cellular respiration which ultimately generates fuel for the cell's activities.

1.1. Evolutionary origin of mitochondria

Unlike any other organelle, except for chloroplasts, mitochondria appear to originate only from other mitochondria. They contain their own DNA, which is circular as is true with bacteria, along with their own transcriptional and translational machinery. Mitochondrial ribosomes and transfer RNA molecules are similar to those of bacteria, as are omponents of their membrane. These and related observations led Dr. Lynn Margulis, in the 1970s, to propose an extracellular origin for mitochondria. Some species of present day protists contain living organisms within their cytoplasm. For example, Paramecium bursaria are hosts for zoochlorellae, photosynthetic protists that reside within the cytoplasm. The relationship appears to be symbiotic. The endosymbiont gains protection and possibly some essential nutrients from the host cytoplasm. The host has a readily available food source when its usual food source is depleted. If you have the opportunity to observe P. bursaria, note that the endosymbionts are not incorporated into food vacuoles. They are residents within the cytoplasm itself, and either is descended from organisms that survived endocytosis or have some mechanism for escaping food vacuoles once they are ingested. Protists are eukaryotes, of course, meaning that their genetic material is organized into a compartment, the nucleus, that is surrounded by membrane, and that they have membrane-delineated organelles. In the warm seas of the ancient earth, the first living things would have been prokaryotes. The endosymbiotic hypothesis for the origin of mitochondria (and chloroplasts) suggests that mitochondria are descended from specialized bacteria (probably purple nonsulfur bacteria) that somehow survived endocytosis by another species of prokaryote or some other cell type, and became incorporated into the cytoplasm. The ability of symbiont bacteria to conduct cellular

respiration in host cells that relied on glycosis and fermentation would have provided a considerable evolutionary advantage. Similarly, host cells with symbiont bacteria capable of photosynthesis would also have an advantage. In both cases, the number of environments in which the cells could survive would have been greatly expanded. Mitochondria do not contain anywhere near the amount of DNA needed to code for all mitochondria-specific proteins, however, a billion or so years of evolution could account for a progressive loss of independence. The endosymbiotic hypothesis might be called a theory, but experimental evidence can't be provided to test it. Only circumstantial evidence is available in support of the proposal, which is the most likely explanation for the origin of mitochondria. The evidence needed to change the model from hypothesis to theory is likely forever lost in antiquity.

1.2. Structure of mitochondria

MtDNA is an extranuclear genome of eukaryotic cells. Each mitochondrion contains usually 5–10 mtDNAs in its matrix. The number of mtDNAs in the somatic cell is about 1,000–10,000 (Lightowlers *et al.* 1997). Human mtDNA is a circular double-stranded molecule with a size of approximately 16.6 kbp (Andrews *et al.* 1999). It codes for 22 distinct transfer RNAs, two ribosomal RNAs and for 13 protein genes, which are mostly involved in the electron transport and oxidative phosphorylation (OXPHOS) pathway — the energy producing system of the cell located within the mitochondrial inner membrane. This system is made up and controlled both by the products of genes, encoded in nuclear DNA and mtDNA. The majority of proteins of OXPHOS enzyme complexes are encoded in the cell nucleus and transported to mitochondria from cytosol. Most of the genes are transcribed from guanine-rich heavy (H) strand of mtDNA, only the sixth subunit of complex I and eight tRNAs are transcribed from cytosine-rich light (L) strand (Wallace *et al.* 1999).

1.3. Energy production by mitochondria

Only a small fraction of the total free energy potentially available from glucose is released in glycolysis. The metabolism of carbohydrates is completed in the mitochondria when pyruvate is imported and oxidized by molecular oxygen (O2) to CO2 and water. The energy released is harnessed so efficiently that about 30 molecules of ATP are

produced for each molecule of glucose oxidized, whereas only 2 molecules of ATP are produced by glycolysis alone. Oxidative metabolism in mitochondria is fuelled not only by pyruvate produced from carbohydrates by glycolysis in the cytosol but also by fatty acids. Pyruvate and fatty acids (from triglycerides) are selectively transported from the cytosol into the mitochondrial matrix, where they are broken down into the two-carbon acetyl group on acetyl coenzyme A (acetyl CoA) by the pyruvate dehydrogenase complex and the ®-oxidation pathway, respectively. The acetyl group is then fed into the tricarboxylic acid cycle for further degradation, and the process ends with the passage of acetyl-derived high-energy electrons along the respiratory chain.

The proteins involved in OXPHOS are located within the mitochondrial inner membrane and include the electron transport chain (ETC) components (complexes I to IV), FOF1 ATP synthase and the adenine nucleotide translocator (ANT). Highenergy electrons from TCA are combined with molecular oxygen by means of the ETC to generate water. These electrons, borne on NADH (nicotinamide adenine dinucleotide), are transferred to respiratory complex I (NADH dehydrogenase) and then to coenzyme Q10 (CoQ), while the electrons from succinate are transferred to complex II (succinate dehydrogenase, SDH) and CoQ. From CoQ, they are passed to complex III, and then to cytochrome c (cyt c), complex IV (cytochrome c oxidase, COX) and finally to ½ O2 to give H2O. The energy released is used to pump protons (H+) out of the mitochondrial matrix, creating an electrochemical gradient across the inner membrane that is positive and acidic on the outside and negative and alkaline on the mitochondrial matrix side. This gradient creates a capacitor that can be depolarized by the transport of protons back into the matrix through a proton channel in the FO membrane component of ATP synthase. This proton flux drives the condensation of adenosine diphosphate (ADP) and inorganic phosphate (Pi) to make ATP, which is then exported to the cytosol in exchange for ADP by the ANT (Figure 1). In this way oxygen consumption by the ETC is coupled to ADP phosphorylation by ATP synthase through the electrochemical gradient (DiMauro & Bonilla 1997, Wallace 1997).

1.4. Inheritance of mitochondria and mtDNA

Female inheritance:

In sexually reproducing organisms, mitochondria are normally inherited exclusively from the mother. The mitochondria in mammalian sperm are usually destroyed by the egg cell after fertilization. Also, most mitochondria are present at the base of the sperm's tail, which is used for propelling the sperm cells. Sometimes the tail is lost during fertilization.The fact that mitochondrial DNA is maternally inherited enables researchers to trace maternal lineage far back in time. (Y chromosomal DNA, paternally inherited, is used in an analogous way to trace the agnate lineage.) This is accomplished in humans by sequencing one or more of the hypervariable control regions (HVR1 or HVR2) of the mitochondrial DNA. HVR1 consists of about 440 base pairs. These 440 base pairs are then compared to the control regions of other individuals (either specific people or subjects in a database) to determine maternal lineage. Most often, the comparison is made to the revised. Vilà *et al* have published studies tracing the matrilineal descent of domestic dogs to wolves. The concept of the Mitochondrial Eve is based on the same type of analysis, attempting to discover the origin of humanity by tracking the lineage back in time. Because mtDNA is not highly conserved and has a rapid mutation rate, it is useful for studying the evolutionary relationships - phylogeny - of organisms. Biologists can determine and then compare mtDNA sequences among different species and use the comparisons to build an evolutionary tree for the species examined.

Male inheritance:

It has been reported that mitochondria can occasionally be inherited from the father in some species such as mussels. Paternally inherited mitochondria have also been reported in some insects such as the fruit fly and the honeybee. Evidence supports rare instances of male mitochondrial inheritance in some mammals as well. While many of these cases involve cloned embryos or subsequent rejection of the paternal mitochondria, others document *in vivo* inheritance and persistence under lab conditions (Hoeh *et al.* 1991).

2. Biology of mtDNA

mtDNA encodes polypeptides involved in electron transport (Attardi and Schatz, 1988) and is maternally inherited. Unlike nuclear DNA, in which there are only two copies of each gene per cell, thousands of copies of mtDNA are present in every nucleated cell. Normal individuals are homoplasmic (that is, virtually all of their mtDNAs are identical). Heteroplasmy (the presence of both normal and mutant mtDNA in a single individual) is present in most mtDNA diseases, so that the proportion of mutant mtDNA in any cell or tissue may range from 0% to 100%.The polypeptides encoded by mtDNA are all subunits of the mitochondrial respiratory chain. This is a highly complex array of multimeric enzymes that generate ATP. Most of these are encoded in the nucleus, as are many mitochondrial proteins.

2. 1. The mitochondrial genome

Mitochondria have been essential for the evolution of animals. It is generally believed that the energy-converting organelles of eucaryotes evolved from procaryotes that were engulfed by primitive eucaryotic cells and developed a symbiotic relationship with them about 1.5 x 109 years ago. This would explain why mitochondria have their own DNA (mtDNA), which codes for some of their proteins. Since their initial uptake by a host cell, these organelles have lost much of their own genome and have become heavily dependent on genes in the nucleus.Extant mammalian mtDNAs have retained only 13 polypeptide genes, all of which encode essential components of OXPHOS. MtDNA also encodes the 12S and 16S rRNA genes and the 22 tRNA genes required for mitochondrial protein synthesis. The remaining mitochondrial OXPHOS proteins, the metabolic enzymes, the DNA and RNA polymerases, the ribosomal proteins and the mtDNA regulatory factors are all encoded by nuclear genes, synthesized in the cytosol and then imported into the organelle (Shoffner & Wallace 1995, Wallace *et al.* 1997a). The transfer of mtDNA sequences to the nucleus is a continuous process (Wallace 1997, Hirano *et al.* 1997), but not all of the mtDNA genes transferred to the nucleus are functional. The presence of hundreds of mtDNA-like sequences or pseudogenes in the human nuclear genome has been well documented (Shay & Werbin 1992). Furthermore, the genetic code in human mitochondria has come to differ from that used in the nucleus, and thus mtDNA genes are no longer intelligible to the nucleocytosolic system (Wallace 1982). UGA is read as

tryptophan rather than 'stop', AGA and AGG as 'stop' rather than arginine, AUA as methionine rather than isoleucine, and AUA or AUU is sometimes used as an initiation codon instead of AUG (Anderson *et al.* 1981, Montoya *et al.* 1981).

2.1.1. Organization of the human mitochondrial genome

The human mitochondrial genome is 16,569 base pairs (bp) in length (Anderson *et al.* 1981), a closed, circular molecule located within the mitochondrial matrix and present in thousands of copies per cell. Mitochondrial DNA has two strands, a guaninerich heavy (H) strand and a cytosine-rich light (L) strand. The heavy strand contains 12 of the 13 polypeptide-encoding genes, 14 of the 22 tRNA-encoding genes and both rRNAencoding genes. Introns are absent in mtDNA, and all of the coding sequences are contiguous (Anderson *et al.* 1981, Wallace *et al.* 1992, Zeviani *et al.* 1998). The only non-coding segment of mtDNA is the displacement loop (D-loop), a region of 1121 bp that contains the origin of replication of the H-strand (OH) and the promoters for L and H-strand transcription. The mtDNA is replicated from two origins. DNA replication is initiated at OH using an RNA primer generated from the L-strand transcript. H-strand synthesis proceeds two-thirds of the way around the mtDNA, displacing the parental Hstrand until it reaches the L-strand origin (OL), situated in a cluster of five tRNA genes. Once exposed on the displaced H-strand, OL folds a stem-loop structure and L-strand 17 synthesis is initiated and proceeds back along the H-strand template. Consequently, mtDNA replication is bidirectional but asynchronous (Clayton 1982). MtDNA transcription is initiated from two promoters in the D-loop, PL and PH. Transcription from both promoters proceeds around the mtDNA circle, creating a polycistronic RNA. The tRNA genes which punctuate the larger rRNA and mRNA sequences then fold within

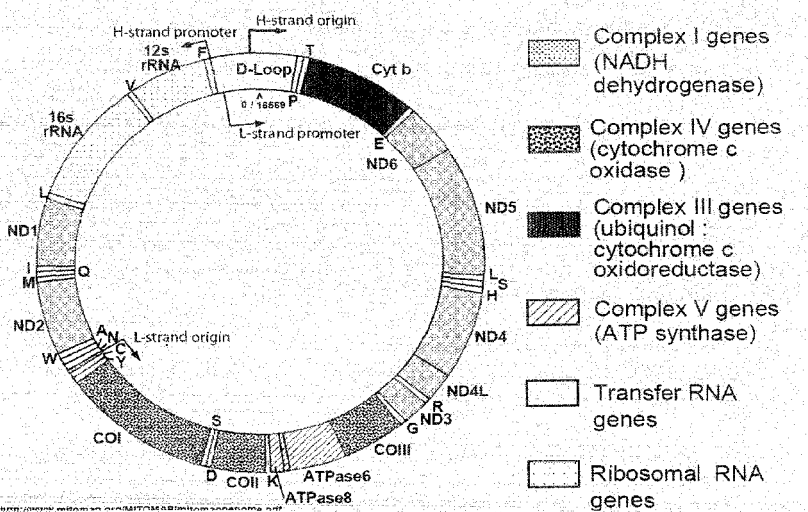


Fig 1 Human mtDNA genomic map (modified from www.mitomap.org).

the transcript and are cleaved out. The mRNAs and rRNAs liberated are posttranscriptionally polyadenylated and the tRNAs are modified and the 3' terminal CCA added (Attardi & Montoya 1983, Clayton 1984, Wallace 1993, Taanman 1999).

2.2. Special features of mitochondrial genetics

The cytoplasmic location of mtDNA and the high copy number contribute to certain unique features of mitochondrial genetics. First, mtDNA is maternally inherited. Second, mtDNA genes have a much higher mutation rate than nuclear DNA genes. Third, mitochondria undergo replicative segregation at cell division. Fourth, many of the pathogenic mtDNA mutations are heteroplasmic. For expression of a disease it is required that a certain threshold level of mutant mtDNA should be exceeded. Fifth, somatic mtDNA mutations accumulate in post-mitotic tissues with age, reducing the ATP generating capacity.

2.2.1. Maternal inheritance and the lack of recombination; mtDNA as a single locus

MtDNA is maternally inherited. The mammalian egg contains about 100,000 molecules of mtDNA, while the sperm contains of the order of 100–1500 mtDNAs (Manfredi *et al.* 1997). Sperm mitochondria enter the egg during fertilization but they appear to be lost early in embryogenesis, soon after fertilization, between the twocell and four-cell stages (Manfredi *et al.* 1997). Paternal mtDNA could not be detected in human neonates born after in vitro fertilization by intracytoplasmic sperm injection (Danan *et al.* 1999). This could be due either to destruction of sperm mitochondria or to impaired replication of sperm mtDNA in the cells (Manfredi *et al.* 1997). However, the presence of paternal mtDNA has been shown at the blastocyst stage in some abnormal (polyploid) human embryos generated by in vitro fertilization and intracytoplasmic sperm injection techniques (St John *et al.* 2000). Due to the lack of recombination, mtDNA acts as a single locus. The effective population size of the mitochondrial genome is only one fourth that of the autosomal loci. The influence of genetic drift thus makes the mitochondrial genome more sensitive to random fluctuations of allele frequencies than that for the autosomal loci.

2.2.2. Homoplasmy and heteroplasmy

Various tissues of the same individual usually share only one type of mtDNA — this condition is referred to as homoplasmy. When a mutation arises, there may be complete

switching to the new mtDNA variant within a single generation (Poulton *et al.* 1998). If the switching is incomplete then two or more types of mtDNA can be observed in a cell. This state is called heteroplasmy. Homoplasmy is thought to be important for the maintenance of the normal mitochondrial function through the coordinated expression of the mitochondrial and nuclear genes (Hirata *et al.* 2002) and is believed to be preserved by the genetic bottleneck mechanism in the oogenesis, most likely in the stage of primordial germ cells. There is no strict definition for the mtDNA bottleneck — it has been described as an event or series of events at one or several stages of oogenesis, which leads to a reduction in the number of mtDNAs, followed by a rapid increase in mature oocytes (Thorburn and Dahl 2001). Different numbers (from 1 to 200) of segregating mtDNA units have been suggested for mammals (Koehler *et al.* 1991; Marchington *et al.* 1998; Poulton *et al.* 1998).

2.2.3. Mutation rate in human mtDNA

On average, mtDNA accumulates mutations more than 10 times faster than does the nuclear genome (Brown *et al.* 1979; Ingman and Gyllensten 2001). This phenomenon leads to high level of within-population polymorphisms. A high mutation rate of mtDNA is associated with several properties of mtDNA, different from those of nuclear genome, and with the peculiarities of processes taking place in the mitochondrial OXPHOS pathway. Firstly, mtDNA lacks protective proteins like histons. Secondly, mtDNA is exposed to oxidative damage by reactive oxygen radicals — by-products of OXPHOS. Thirdly, it has been suggested that the reparation system of mtDNA is not as effective as that in the nucleus (Bogenhagen 1999), although many elements of the mtDNA repair system have been shown to exist in animal mitochondria (Mason and Lightowlers 2003). The fast evolving mtDNA provides more information about recent events in evolution than does a strech of DNA of an equal length in the nuclear genome. At the same time, the high mutation rate creates possibilities for homoplasy — the same mutation could arise in different branches of the mtDNA phylogenetic tree, thus blurring the possibility to establish an unambiguous (more precisely — the most parsimonious) order of evolutionary events. The mutation rate is different both for mtDNA regions and nps within a region. Pesole *et al.* (1999) showed that: 1) nonsynonymous nps, D-loop central domain, tRNA and rRNA genes evolve at about a five to ten times lower rate than

synonymous nps and two peripheral domains of D-loop; 2) the rate of synonymous sites is quite uniform over the genome, whereas the rate of nonsynonymous sites differs considerably between genes; 3) nonsynonymous sites show comparable rates in mtDNA and nuclear genome, synonymous sites and rRNA evolve *ca* 20 times and tRNAs *ca* 100 times more rapidly in mammalian mitochondria. Additionally, transitions have been shown to occur *ca* 12 to 37 times more often than transversions (Meyer *et al.* 1999). The considerable rate variation has also been observed between different Dloop nps (Macaulay *et al.* 1997; Finnilä *et al.* 2001). For example, the transitions at nps 16093, 16129, 16189, 16311 and 16362 in HVS-I and 73, 146, 150, 152, 195 in HVS-II are considered as mutational "hotspots" and are often observed in different phylogenetic branches of mtDNA. It also raises a possibility that, in "hotspots", potential recurrent mutations may be missed or overlooked in phylogenetic reconstructions. In phylogenetic tree-building, differences in mutation rates can be normalized by assigning different weights to the nps with known rate variation (Richards *et al.* 1998). However, "hidden" and/or parallel mutations do not likely harm the outcome in any profound way, provided the level of resolution between the branches of an mtDNA phylogenetic tree is sufficient, i.e. the information of basal nodes of the tree is available. Then, the fast evolving positions may, theoretically, blur only the terminal tips of the tree, not distorting its basic topology. The combined usage of information both from fast evolving control region sequences and diagnostic coding region sites has justified itself in many mtDNA population genetic studies (Torroni *et al.* 1996; Richards *et al.* 1998; 2000; Macaulay *et al.* 1999*b*;Kivisild *et al.* 2002).

2.2.4. Replicative segregation

Each cell has hundreds of mitochondria, each containing 2 to 10 copies of mtDNA molecules. Normally all mtDNAs in a cell are identical, a condition known as homoplasmy. At cell division, the mitochondria and their genomes are randomly distributed to the daughter cells, a process known as replicative segregation. When a mutation arises in mtDNA, it creates an intracellular mixture of mutant and normal molecules, a condition known as heteroplasmy. Despite the high mtDNA copy number in mature oocytes and the relatively small number of cell divisions in the female germline, mtDNA sequence variants segregate rapidly between generations (Poulton *et al.* 1998).

This has been attributed to a genetic bottleneck. A major component of the mtDNA bottleneck occurs by the time that oocytes are mature (Jenuth *et al.* 1996, Marchington *et al.* 1998). Jenuth *et al.* (1996) have estimated that in oogenesis the effective number of segregating units for mtDNA is approximately 200 in mice. The model used by Jenuth *et al.* (1996) assumes that the variance in genotypic ratios of the progeny or developing oocytes is caused by an identical randomsampling event that occurs during each of the 15 or so cell divisions during the later stages of oogenesis (i.e. repeated selection), in contrast to a more dramatic reduction in segregating units during a briefer period (i.e. single selection). On the basis of single selection model it is estimated that the most probable bottleneck size is 1–31 segregating units in humans (Marchington *et al.* 1998). Both estimations of the number of segregating units are far below the number of mtDNA molecules in a cell, suggesting that there is first a restriction in the quantity of mtDNA to be transmitted, followed by amplification, and thereby constituting a genetic bottleneck (Poulton *et al.* 1998).

2.2.5. Neutrality

Neutrality of the markers being used is a necessary precondition for many methods in phylogenetic analysis. Several tests of neutrality have been performed, measuring the difference between the fixation of non-synonymous and synonymous mutations in mtDNA at intra- and interspecies level (Excoffier 1990; Graven *et al.* 1995; Nachman 1998; Nielsen and Weinreich 1999; Whittam *et al.* 1986). Torroni *et al.* (1994) showed that mtDNA mutations are unlikely to be related to the Tibetans' adaptation to living in high altitudes. Those nonsynonymous substitutions that are known to be under selection are normally excluded from phylogenetic analyses. Several substitutions have been found in association with the inheritance of diseases (Wallace *et al.* 1999). Mildly deleterious substitutions under selection should get eliminated in long run. However, when a mildly deleterious substitution occurs in a small expanding population, it can, on the contrary, be maintained (Excoffier 1990). A good example for the latter case could be haplogroup J in Europe. Several mutations specific for haplogroup J, which has probably expanded in Europe together with the Neolithic spread of farmers (Richards *et al.* 1996), have been suggested to play a background role in LHON expression (Torroni *et al.* 1997). Otherwise, the frequency of neutral substitutions when linked with a substitution under

selection can be shifted up or down by hitch-hiking (Smith and Haigh 1974), as far as mtDNA is a single locus. Yet, being as well the subject of the cumulative chance fixation of mildly deleterious mutations (because of the lack of recombination and clonal inheritance), mitochondria are likely to be led to Muller's ratchet (Lynch 1996). The properties of mtDNA mentioned above, with the given restrictions, allow one to reconstruct genealogies of individuals through maternal descent. This option for 'trees of individuals' makes mitochondria different from markers whose variation is expressed only in allele frequencies and/or evolve too slowly for revealing genealogies through their mutational pattern.

2.2.6. The role of natural selection in the evolution of human mtDNA

The assumption of the neutrality of a genetic marker is one of the cornerstones of many methods that are used in phylogenetic analysis. The theory of neutral evolution was first introduced by Kimura (1968). According to this theory, mutations occur stochastically and their fixation is the result of random drift rather than of natural selection. Occurring deleterious mutations are removed by purifying selection; positive selection does not play any significant role. This means, theoretically, that the rate of evolution solely depends on the mutation rate. Correspondingly, the simplest model for explaining the present mtDNA variation is the following: mutations have accumulated sequentially along radiating female lineages and have reached to polymorphic frequencies only because of random genetic drift in its various manifestations, whereas the influence of (positive) natural selection has been negligible. The demographic history of a population, or more precisely — of carriers of particular variants — has likely played a decisive role. According to the "near-to-neutral" theory of evolution there may also be, among the mutations of recent origin in the evolutionary time scale, slightly deleterious ones that are not yet removed by purifying selection (Hasegawa *et al.* 1998; Nachman 1998; Gerber *et al.* 2001). The lack of recombination in human mtDNA makes it a subject of "Muller's rachet" — of a genetic mechanism that predicts an accumulation of slightly deleterious mutations. Many authors have tested the neutrality of mtDNA evolution, primarily by estimating the differences between the fixation of nonsynonymous and synonymous substitutions in the mtDNA of different species. They have observed that there is an excess of nonsynonymous mtDNA polymorphisms relative to fixed sequence change (Graven *et al.*

1995; Nachman 1998; Excoffier and Yang 1999). It has been also shown that the "older" branches of the human mtDNA tree contain relatively less nonsynonymous substitutions than the "younger" ones, which can be interpreted by inferring the action of purifying selection on the mtDNA (Elson *et al.*2004; Ruiz-Pesini *et al.* 2004). Recent studies of human populations have compared the ratios of nonsynonymous and synonymous substitutions in different lineages of complete mtDNA sequences from Africa, Asia and Europe (Mishmar *et al.* 2003; Elson *et al.* 2004; Ruiz-Pesini *et al.* 2004). Mishmar *et al.* (2003) and Ruiz-Pesini *et al.* (2004) interpreted the differences between the results of different haplogroups spread in tropical, temperate and arctic zones as evidence for climatic adaptations, and noted that the present continental distribution of mtDNA haplogroups might be non-random, shaped by positive selection. The analysis, however, did not consider the possibility that analogous differences might also exist between lineages that have not been selected according to their geographical origin, as it has been demonstrated (Excoffier 1990; Torroni *et al.* 2001; Moilanen and Majamaa 2003; Elson *et al.* 2004). It has to be noted that the neighbour-joining tree used by Ruiz-Pezini and coauthors was incorrect since the branching order of its clusters did not correspond to the established phylogenetic relationships between the clades of mtDNA. Due to these special properties, mtDNA offers an opportunity to reconstruct the maternal genealogies, unhindered by the genetic fog of recombination and to detect the differences even among closely related groups that have diverged within a relatively short time scale. Thus, the genealogy of maternal lineages provides a link between observable sequence variation and evolutionary events that have shaped this diversity. However, one should not forget that in order to draw conclusions about the demographic history of a population the data of different genetic systems (Y chromosome, autosomes) must be analysed and combined with those obtained from maternally inherited mtDNA.

2.2.7. Threshold effect

Many but not all pathogenic mtDNA mutations are heteroplasmic. The phenotype is normal until a critical proportion of mutant mtDNA is present within the tissue and the threshold for genotype expression is exceeded (Wallace *et al.* 1997a). This threshold varies for different types of mtDNA mutation and is about 60% for deleted mtDNA (Hayashi *et al.* 1991). For the mutation 8344A>G, which causes the syndrome of

myoclonic epilepsy and ragged-red fibers, the threshold level is about 85% mutated DNA (Chomyn 1998). Once this is exceeded, large changes in the phenotype can be observed with minor increases in the proportion of the mutant mtDNA. Different phenotypes associated with the same genotype are determined mainly by the localized concentration and distribution of the mutation in affected tissues (Petruzzella *et al.* 1994). Furthermore, different tissues have different dependences on oxidative phosphorylation for normal function. Organs with the highest ATP requirements and the lowest regenerative capacities, such as the brain, heart and skeletal muscle, are the most sensitive to the effects of pathogenic mtDNA mutations (Wallace 1994, 1995).

2.2.8. mtDNA genetic bottleneck

The research has proven that there is a "mitochondrial genetic bottleneck", where only a small number of mtDNA molecules in the mother are passed on to the next generation.

All human and animal cells contain many mitochondria, responsible for energy production in the body. These mitochondria have their own genetic information, known as mitochondrial DNA, or mtDNA, which is inherited. Although, a child inherits one copy of DNA from each parent, a child may inherit many copies of mtDNA, which are only passed down from the mother.Mutations in mtDNA can affect energy production within cells and thus causes the disease. Though, mitochondrial diseases differ both in location and severity depending on where and at what levels the mutations are distributed, defective mitochondria leads to maximum damage in muscles, nerves and the brain, as these are the parts of the body consuming most of the energy (http://www.medindia.net/news/).Rapid intergenerational changes in mitochondrial genotype were first observed in Holstein cows transmitting mtDNA polymorphisms (Upholt and Dawid 1977; Hauswirth and Laipis 1982; Olivo *et al.* 1983). Similar results were subsequently described in many mammalian species, including humans transmitting pathogenic mtDNA mutations (Holt et al. 1989; Vilkki *et al.* 1990; Larsson *et al.* 1992). These observations led to the suggestion that only a small number of mtDNA molecules were passed on from mother to offspring—the mitochondrial 'genetic bottleneck' (Hauswirth and Laipis 1982).

2.2.9. Age-related somatic mtDNA mutations

Oxygen free radicals damage mtDNA, causing oxidative modifications of DNA bases,

base substitutions and rearrangements. The cumulative accumulation of these somatic mutations during life may cause a bioenergetic deficit leading to cell death, or apoptosis, and normal ageing (Ozawa 1995). In addition to the ageing or senescence process somatic mtDNA mutations may be important for determining the onset and progression of mtDNA diseases. Most inherited mutations are insufficient to suppress mitochondrial OXPHOS below the expression threshold and thus it is the accumulation of somatic mutations in postmitotic tissues that exacerbates the inherited OXPHOS defect and ultimately leads to phenotypic expression (Wallace 1995). Oxidative stress has been implicated in the pathogenesis of neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease. There is much evidence of increased oxidative stress and free radical damage in the substantia nigra in patients with Parkinson's disease, and there is also evidence for a defect in mitochondrial energy production, and especially reduced complex I activity, in the substantia nigra (Schapira 1990).

2.2.10. mtDNA mutation and adaptation

The potential for temperature to influence the evolution of mtDNA was the focus of much attention a decade ago (Martin & Palumbi 1993, Rand1994). Differences in mtDNA substitution rates suggest that the thermal environment may influence rates of evolution and indicate that it is inappropriate to use a calibration for one group to estimate divergence times or demographic parameters for another group. It has also been argued that thermal adaptation may occur within humans. Ruiz-Pesini *et al.* (2004) conducted a phylogenetic study including 1125 globally distributed human mtDNA sequences and observed that the relative frequency and amino acid conservation of internal branch amino acid mutations increased from tropical Africa to temperate Europe to arctic northeastern Siberia. Highly conserved amino acid changes were found at the roots of multiple mtDNA lineages from higher latitudes prompting the authors to suggest that specific mtDNA nonsynonymous mutations permitted our ancestors to adapt to more northern climates because their mitochondria produced more heat. However, Elson *et al.* (2004) analyzed complete mtDNA coding-region sequences for 560 maternally unrelated individuals of European, African, and Asian descent and were not able to replicate the results of Ruiz-Pesini *et al.* (2004). One clear alternative is that the mtDNA is under strong thermal selection associated with the different temperature between Atlantic and

Gulf waters. However, in no case was the disjunction in the same geographic region as that observed for the mtDNA. Ropson et al. consider a variety of alternatives "in the absence of selection" but never fully explore the possibility that selection may operate on the mtDNA itself.

2.2.11. The role of natural selection in the evolution of human mtDNA

The assumption of the neutrality of a genetic marker is one of the cornerstones of many methods that are used in phylogenetic analysis. The theory of neutral evolution was first introduced by Kimura (1968). According to this theory, mutations occur stochastically and their fixation is the result of random drift rather than of natural selection. Occurring deleterious mutations are removed by purifying selection; positive selection does not play any significant role. This means, theoretically, that the rate of evolution solely depends on the mutation rate. Correspondingly, the simplest model for explaining the present mtDNA variation is the following: mutations have accumulated sequentially along radiating female lineages and have reached to polymorphic frequencies only because of random genetic drift in its various manifestations, whereas the influence of (positive) natural selection has been negligible. The demographic history of a population, or more precisely — of carriers of particular variants — has likely played a decisive role. According to the "near-to-neutral" theory of evolution there may also be, among the mutations of recent origin in the evolutionary time scale, slightly deleterious ones that are not yet removed by purifying selection (Hasegawa *et al.* 1998; Nachman 1998; Gerber *et al.* 2001). The lack of recombination in human mtDNA makes it a subject of "Muller's rachet" — of a genetic mechanism that predicts an accumulation of slightly deleterious mutations. Many authors have tested the neutrality of mtDNA evolution, primarily by estimating the differences between the fixation of nonsynonymous and synonymous substitutions in the mtDNA of different species. They have observed that there is an excess of nonsynonymous mtDNA polymorphisms relative to fixed sequence change (Graven *et al.* 1995; Nachman 1998; Excoffier and Yang 1999). It has been also shown that the "older" branches of the human mtDNA tree contain relatively less nonsynonymous substitutions than the "younger" ones, which can be interpreted by inferring the action of purifying selection on the mtDNA (Elson *et al.*2004; Ruiz-Pesini *et al.* 2004). Recent studies of

human populations have compared the ratios of nonsynonymous and synonymous substitutions in different lineages of complete mtDNA sequences from Africa, Asia and Europe (Mishmar *et al.* 2003; Elson *et al.* 2004; Ruiz-Pesini *et al.* 2004). Mishmar *et al.* (2003) and Ruiz-Pesini *et al.* (2004) interpreted the differences between the results of different haplogroups spread in tropical, temperate and arctic zones as evidence for climatic adaptations, and noted that the present continental distribution of mtDNA haplogroups might be non-random, shaped by positive selection. The analysis, however, did not consider the possibility that analogous differences might also exist between lineages that have not been selected according to their geographical origin, as it has been demonstrated (Excoffier 1990; Torroni *et al.* 2001; Moilanen and Majamaa 2003; Elson *et al.* 2004). It has to be noted that the neighbour-joining tree used by Ruiz-Pezini and coauthors was incorrect since the branching order of its clusters did not correspond to the established phylogenetic relationships between the clades of mtDNA. Due to these special properties, mtDNA offers an opportunity to reconstruct the maternal genealogies, unhindered by the genetic fog of recombination and to detect the differences even among closely related groups that have diverged within a relatively short time scale. Thus, the genealogy of maternal lineages provides a link between observable sequence variation and evolutionary events that have shaped this diversity. However, one should not forget that in order to draw conclusions about the demographic history of a population the data of different genetic systems (Y chromosome, autosomes) must be analysed and combined with those obtained from maternally inherited mtDNA.

2.2.12. mtDNA and speciation

Many researchers have assumed explicitly or implicitly that differentiation within a character system is indicative of organismal differentiation or history. From the standpoint of mtDNA, attention has been focused on whether mtDNA differentiation is indicative of species trees or gene trees. A genealogical species is defined as a basal group of organisms whose members are all more closely related to each other than they are to any organisms outside the group and which contains no exclusive group within it (Baum & Shaw 1995; Cracraft 1983). In practice, a pair of species is so defined when phylogenies of alleles from a sample of loci show them to be reciprocally monophyletic at all, or monyphyletic for some specified fraction of the loci (though not all genealogical

species concepts require reciprocal monophyly). Hudson & Coyne (2002) investigated the length of time it takes to attain reciprocal monophyly when an ancestral population divides into two descendant populations of equal size with no gene exchange and when genetic drift and mutation are the only operating evolutionary forces. Funk & Omland (2003) also argue that data from mtDNA may not provide an accurate measure of species status. To evaluate the importance of species-level polyphyly, Funk & Omland (2003) conducted an intense survey of studies that evaluate mtDNA variation in a phylogenetic context and observed that the mtDNA monophyly of many biological species is not well supported. Funk & Omland (2003) detected species-level mtDNA paraphyly or polyphyly in 23% of 2319 assayed species, demonstrating that this problem is statistically supported, taxonomically widespread, and far more common than previously recognized. These patterns could be due to (*a*) sporadic hybridization among divergent lineages, (*b*) incomplete lineage sorting of mtDNA gene trees relative to organismal lineages, and/or (*c*) selection on mtDNA that might retain certain haplotypes within diverging lineages. Moreover, mtDNA has great potential for becoming monophyletic by selective sweeps. This can decrease the time to monophyly of a clade and not be reflective of the genealogical processes in the nuclear genome. Advantageous mutations occurring on mtDNA will cause the entire organelle genome to become monophyletic because such genomes have little or no recombination. Although selective sweeps will also occur in nuclear DNA, causing monophyly for regions linked to the selected locus, recombination will whittle away the section of genome that becomes monophyletic through linkage. Finally, when there is gene flow between diverging populations, one may encounter the opposite problem: mtDNA may be homogenized between the populations more readily than is nuclear DNA, so that mtDNA may appear paraphyletic when nuclear genes may be monophyletic. In fish, mice, and crickets, for example, mtDNA flows between taxa much more readily than does nuclear DNA (Shaw 2002, Taylor & McPhail 2000). In some cases, the mtDNA from one taxon completely replaces that in another, without any evidence of nuclear introgression or morphological signal.

3. Mitochondrial DNA sequence variation in human populations

The average number of base pair differences between two human mitochondrial genomes is estimated to be from 9.5 to 66 (Zeviani *et al.* 1998). The high mutation rate has

resulted in the accumulation of a wide range of neutral, population-specific base substitutions in mtDNA. These have accumulated sequentially along radiating maternal lineages that have diverged approximately on the same time scale as human populations have colonized different geographical regions of the world. Thus the women that migrated out of Africa into the different continents about 130,000 years before present (YBP) harboured mtDNA mutations which today are seen as high-frequency, populationspecific mtDNA polymorphisms creating groups of related mtDNA haplotypes, or haplogroups (Torroni & Wallace 1994, Wallace 1995).

4. Mitochondrial DNA haplogroups

MtDNA haplogroups determined by polymorphisms that occurred tens of thousands of years ago are today high-prevalence population-specific substitutions. Haplotypes are subclusters of haplogroups, and the polymorphisms that determine them are less prevalent and have occurred more recently. Most of the polymorphisms determining haplogroups are continent-specific (Wallace 1994). The classification of mtDNA haplogroups is based on information gained from RFLP analysis of the coding region and from the nucleotide (nt) sequences of the hypervariable segments (HVS) in the control region. Haplogroups are coded with capital letters and subclusters with a running number (Schurr *et al.* 1990, Ballinger *et al.* 1992a, Torroni *et al.* 1996, Richards *et al.* 1998.) MtDNA has been used for a couple of decades as a molecular marker in population genetics. The earliest work began with digestion of the molecule, either with a single restriction enzyme in a large number of samples (Denaro *et al.* 1981), or with many enzymes in a few samples (Brown 1980), and subsequent studies tended to use five or six enzymes on fairly large sample sets . 'High-resolution' restriction analysis with 14 enzymes enabled a much more detailed mtDNA phylogeny to be obtained (Cann *et al.* 1987). Analysis using 14 restriction endonucleases allowed screening of 15-20% of the mtDNA sequence for variations (Chen *et al.* 1995b), and a further improvement in resolution was obtained when information on the control region sequence was included with the RFLP data (Macaulay *et al.* 1999). The best resolution that can be obtained is gained by analysing complete mtDNA sequences which have begun to emerge (Richards & Macaylay 2001).

4.1. How to classify mtDNA haplogroups

All humans living today can be classified into a specific mtDNA haplogroup in the mtDNA Tree based on SNP markers found in their mtDNA. The mtDNA consists of 3 main regions, namely HVR-1, HVR-2 and Coding region. SNP markers can be found in all 3 regions of the mtDNA. Testing SNPs the HVR-1 and HVR-2 regions of the mtDNA allows scientists to predict an individual's haplogroup. However, testing all 3 regions of the mtDNA is usually required in order to confirm an individual's haplogroup and subclade. Human mtDNA haplogroups are named using the letters of the alphabet, A to Z. Each mtDNA haplogroup can be further subclassified into finer branches called subclades. Subclades are named using numbers and lower case letters. For example, H1a is a subclade of mtDNA haplogroup H. Mitochondrial Eve is the name given by researchers to the woman who is the most recent common matrilineal ancestor of all living humans and at the root of the mtDNA phylogenetic tree. **(http://www.dnahaplogroups.org/)**

4.2. Mitochondrial DNA haplogroups on different continents

Human mtDNA haplogroups are lettered: A, B, C, CZ, D, E, F, G, H, pre-HV, HV, I, J, pre-JT, JT, K, L0, L1, L2, L3, L4, L5, L6, L7, M, N, P, Q, R, S, T, U, UK, V, W, X, Y, and Z. The letter names of the haplogroups run from A to Z. As haplogroups were named in the order of their discovery, they do not reflect the actual genetic relationships.

The common divisions for mtDNA haplogroups:

- Sub-Saharan African: L0, L1, L2, L3, L4, L5, L6, L7
- West Eurasian: H, T, U, V, X, K, I, J, W (all listed West Eurasian haplogroups are derived from macro-haplogroup N)
- East Eurasian: A, B, C, D, E, F, G, Y (note: C, D, E, and G belong to macro-haplogroup M)
- Native American: A, B, C, D, X
- Australo-Melanesian: P, Q, S

The mitochondrial haplogroups are divided into 3 main groups, which are designated by the 3 sequential letters L, M, N. Humanity first split within the L group between L0 and L1. L1 gave rise to other L groups, one of which, L3, split into the M and N group. The

M group comprises the first wave of human migration out of Africa, following an eastward route along southern coastal areas. Descendent populations belonging to haplogroup M are found throughout East Africa, Asia, the Americas, and Melanesia, though almost none have been found in Europe. The N group may represent another migration out of Africa, heading northward instead of eastward. Shortly after the migration, the large R group split off from the N. Haplogroup R consists of two subgroups defined on the basis of their geographical distributions, one found in southeastern Asia and Oceania and the other containing almost all of the modern European populations. Haplogroup N (xR), i.e. mtDNA that belongs to the N group but not to its R subgroup, is typical of Australian aboriginal populations, while also being present at low frequencies among many populations of Eurasia and the Americas.

The L type consists of Africans, and especially Sub-Saharan Africans.

The M type consists of:

M1- Ethiopian, Somali and Indian populations. Likely due to much gene flow between the Horn of Africa and the Arabian Peninsula (Saudi Arabia, Yemen, Oman), separated only by a narrow strait between the Red Sea and the Gulf of Aden.

CZ- Many Siberians; branch C- Some Amerindian; branch Z- Many Saami, some Korean, some North Chinese, some Central Asian populations.

D- Some Amerindians, many Siberians and northern East Asians

E- Malay, Borneo, Philippines, Taiwan aborigines, Papua New Guinea

G- Many Northeast Siberians, northern East Asians, and Central Asians

Q- Melanesian, Polynesian, New Guinean populations

The N type consists of:

A- Found in some Amerindians, Japanese, and Koreans

I- 10% frequency in Northern, Eastern Europe

S- Some Australian aborigines

W- Some Eastern Europeans, South Asians, and southern East Asians

X- Some Amerindians, Southern Siberians, Southwest Asians, and Southern Europeans

Y- Ainus and Nivkhs; 1% in Southern Siberia

R- Large group found within the N type.Populations contained therein can be divided geographically into West Eurasia and East Eurasia. Almost all European populations and a large number of Middle-Eastern populations today are contained within this branch. A smaller percentage is contained in other N type groups (See above). Below are **subclades of R**:

B- Some Chinese, Tibetans, Mongolians, Central Asians, Koreans, Amerindians, South Siberians, Japanese, Austronesians

F- Mainly found in southeastern Asia, especially Vietnam; 8.3% in Hvar Island in Croatia.

R0- Found in Arabia and among Ethiopians and Somalis; branch HV (branch H; branch V)- Europe, Western Asia, North Africa;

Pre-JT- Arose in the Levant (modern Lebanon area), found in 25% frequency in Bedouin poupulations; branch JT (branch J; branch T) - North, Eastern Europe, Indus, Mediterranean

U- High frequency in Scandinavia, Baltic countries, Mediterranean
(http://en.wikipedia.org/)

5. Mutations in mtDNA as causes of diseases

The biochemical and genetic complexity of the respiratory chain accounts for the extraordinarily wide range of clinical presentations of mitochondrial disorders. In general, the organs with the highest aerobic demand and the lowest regenerative capacity, such as the brain, heart and skeletal muscle, are the ones that are most severely involved, but virtually any organ or tissue in the body can be affected, including the gastrointestinal tract (Bardosi *et al.* 1987), liver (Mazziotta *et al.* 1992), kidney (Manouvrier *et al.* 1995)

and the endocrine systems (Shoffner *et al.* 1995a, Manouvrier *et al.* 1995). Each tissue can be affected alone, e.g. pure mitochondrial myopathy, encephalopathy or cardiomyopathy, or more often in combination, e.g. mitochondrial encephalomyopathy. The two main biochemical features in most mitochondrial diseases are respiratory chain deficiency and lactic acidosis. Morphologically, patients often display ragged-red fibers (RRF) in the muscle due to the accumulation of structurally abnormal subsarcolemmal mitochondria (Rowland *et al.* 1991). Some mitochondrial syndromes are well established, having a certain molecular genetic background and are nosologically defined entities. There are many disorders, however, that are defined on the basis of morphological or biochemical findings and still lack a molecular genetic definition. In addition, overlap syndromes and non-specific phenotypes mean that clinical data are not sufficient to provide a systematic classification of mitochondrial diseases (Zeviani *et al.* 1996). Genetically, mitochondrial diseases can be divided into three groups: those characterized by the presence of sporadic or maternally inherited mtDNA mutations, those characterized by the association of mtDNA abnormalities with mendelian transmission of the trait, i.e. disorders believed to be due to mutations in nuclear genes that control mitochondrial biogenesis, and those that lack a mtDNA defect but are thought on the basis of biochemical findings to be caused by mutations in nuclear genes.

5.1. Point mutations

Many but not all pathogenic mtDNA point mutations are heteroplasmic. When the proportion of the mutant genome exceeds a certain threshold, the deleterious effects of the mutation will no longer be complemented by the coexisting wild-type mtDNA and will be expressed phenotypically as a cellular dysfunction leading to disease (Wallace *et al.* 1997a). Phenotypic expression will depend on the nature of the mutation, its tissue distribution and the relative reliance of each organ system on the mitochondrial energy supply (Schon *et al.* 1997). The influence of nuclear genes, coexisting mitochondrial polymorphisms, the age and sex of the individual and environmental factors may also play an important, although poorly understood, role in the phenotypic expression of mtDNA point mutations (Wallace *et al.* 1997a). Mitochondrial DNA point mutations are maternally inherited and can occur in rRNA or tRNA genes, or in genes coding for proteins of respiratory chain complexes. Although more than 50 deleterious point

mutations have been identified to date, four mutations are by far the most frequent (Wallace *et al.* 1997a). These are the 243A>G 'MELAS', the 8344A>G 'MERRF', the 8993T>G 'NARP' and the 11778G>A 'LHON' mutations. Others are found less often, while still others have been described only in single individuals or families. The investigation of pathogenic mitochondrial DNA mutations has revealed a complex relation between patient genotype and phenotype (Schon *et al.* 1997). For unknown reasons, some mtDNA mutations lead to specific clinical manifestations, an example being 3243A>G, causing the MELAS syndrome, one of the classic mitochondrial syndromes. Moreover, the MELAS syndrome has a high prevalence in the adult population, suggesting that mitochondrial disorders constitute one of the largest diagnostic categories of neurogenetic diseases (Majamaa *et al.* 1998).

5.2. Large-scale rearrangements of mtDNA

Large-scale rearrangements of mtDNA can be either mtDNA deletions or, more rarely, duplications. Both types of mutation are heteroplasmic, and they can occasionally exist simultaneously in patient tissues (Ballinger *et al.* 1992). More than 120 mtDNA deletions have been identified, most within direct repeats of 3–13 nucleotides in length (Kogelnik *et al.* 1998). The most common large-scale deletions are between a 13-base pair direct repeat from nt 8470 to nt 8482 in the ATPase8 gene and from nt 13447 to nt 13459 in the ND5 gene (Moraes *et al.* 1989). The resulting 4997 bp deletion (the common deletion) has occurred independently over 200 times and accounts for perhaps 50% of ocular myopathy patients (Kogelnik *et al.* 1998). Because virtually all deletions eliminate at least one tRNA, it is likely that they result in a generalized translational defect. Deletions are usually sporadic, non-transmittable mutagenic events (Brown & Wallace 1994). The three main clinical phenotypes associated with large-scale mtDNA deletions are the Kearns-Sayre syndrome, chronic progressive external ophthalmoplegia (CPEO) and the Pearson syndrome (Moraes *et al.* 1989, Poulton *et al.* 1989, Rötig *et al.* 1990, McShane *et al.* 1991). The Kearns-Sayre syndrome is characterized by an invariant triad of PEO, pigmentary retinopathy and onset before 20 years of age. Frequent additional symptoms are a progressive cerebellar syndrome, heart block and increased protein content in the cerebrospinal fluid. CPEO is characterized by bilateral ptosis and ophtalmoplegia, frequently associated with variable degrees of proximal muscle weakness and wasting,

and exercise intolerance. The Pearson bone marrow-pancreas syndrome is a rare disorder of early infancy characterized by sideroblastic anaemia with pancytopaenia and exocrine pancreatic insufficiency. Infants surviving into childhood may develop the clinical features of Kearns-Sayre syndrome (McShane *et al.* 1991, Rötig *et al.* 1990). Partial duplications of mtDNA has been detected in ocular myopathy and Pearson's syndrome patients, although duplications are much rarer in these patients than deletions. Duplications can be sporadic (Poulton *et al.* 1991) or maternally transmitted (Rötig *et al.* 1992).

5.3. MtDNA genotypes conferring increased risk of disease

The high mutation rate has resulted in the accumulation of a wide range of populationspecific base substitutions in mtDNA. While most of these variants are neutral, some are mildly deleterious. The latter, although they do not significantly reduce fitness, may interact with nuclear and environmental factors, predisposing individuals to an increased risk of developing neurodegenerative diseases late in life (Wallace 1994). Moreover, mildly deleterious polymorphisms may synergistically compromise mitochondrial function and contribute to the pathogenesis of a mitochondrial disorder (Lertrit *et al.* 1994). It has been shown that the expression of some pathogenic mtDNA mutations depends on the mtDNA background against which they occur. The risk of expression of LHON in the presence of the primary mutation 14484T>C is eight-fold higher when this mutation occurs in the specifically European haplogroup J (Torroni *et al.* 1996b). The 11778G>A mutation was observed in a wide range of mtDNA haplogroups but has shown an almost six-fold preferential association for haplogroup J (Torroni *et al.* 1997). The mutation 4336T>C is associated with late-onset Alzheimer disease (Hutchin & Cortopassi 1995), and has arisen as a single mutational event in the European haplogroup H (Torroni *et al.* 1994). The association of a particular mtDNA sequence variant with a particular disease is not an unambiguous indicator of aetiological significance (Chinnery *et al.* 1999), however, as the mtDNA sequence may act as a surrogate marker for a nuclear gene defect, particularly in isolated populations that have experienced a marked founder effect (Heyer 1995, Jorde *et al.* 1995). Similarly, a particular mtDNA haplotype may signal, through a founder effect, a population subgroup that has inherited a group of detrimental or protective nuclear genes.

5.4. Defects in nucleo-mitochondrial signalling and nuclear gene defects

Most of the genes involved in mitochondrial functions reside in the nucleus. These include the majority of genes encoding polypeptides for respiratory chain complexes and ATP synthase, and the genes that control the biogenesis of the mitochondrial genome and mitochondrial metabolism. Mitochondrial disorders that still lack molecular genetic characterizations but are defined biochemically as defects of the respiratory chain complexes or are inherited as mendelian traits indicate the existence of mutations in nuclear genes or faulty interactions between nuclear and mitochondrial genes. Virtually any nuclear mutation that alters a pathway in intermediate metabolism, which interfaces with the mitochondria could also interact with OXPHOS gene polymorphisms and lead to disease (DiMauro 1999). The Leigh syndrome is the most common mitochondrial encephalomyopathy of infancy or childhood (Rahman *et al.* 1996), with the characteristic neuropathology of focal, bilaterally symmetrical spongiform lesions, especially in the thalamus and brain stem regions (Dahl 1998). It is much more commonly transmitted as an autosomal recessive trait than as a maternally inherited trait, and is occasionally transmitted as an autosomal dominant or X-linked trait. The Leigh syndrome is one of the most commonly recognized disorders of mitochondrial energy production and can be regarded as one part of the spectrum of these disorders. For example, some mutations in mtDNA, such as 8993T>G or 8993T>C, when present in high proportions in affected tissues, will cause Leigh syndrome, whereas lower levels of mutant mtDNA will cause milder phenotypes (Santorelli *et al.* 1993). In a similar manner, mutations in some nuclear genes appear to result in Leigh syndrome, but functionally less severe mutations in the same genes give rise to different clinical presentations (Dahl 1998). The first pathogenic mutation in nuclear genes was found in two sisters with Leigh syndrome and complex II deficiency (Bourgeron *et al.* 1995). Complex II is entirely encoded by nuclear DNA, and the mutation was detected in the gene encoding the flavoprotein subunit. A few mutations have also been found in nuclear DNA encoding complex I subunits in patients with Leigh syndrome (van den Heuvel *et al.* 1998, Triepels *et al.* 1999), and some have been identified in a nuclear SURF1 gene encoding a protein needed for proper COX assembly (Tiranti *et al.* 1998). Certain abnormalities of mtDNA have been shown to be secondary to nuclear mutations. Autosomal dominant chronic progressive external ophthalmoplegia

is a genetically heterogeneous disorder, with at least three distinct genomic loci causing very similar phenotypes (Kaukonen *et al.* 1996, 1999). All affected patients have multiple mtDNA deletions (Kaukonen *et al.* 1999). Furthermore, two other autosomally inherited diseases with multiple mtDNA deletions, Wolfram syndrome and a recessively inherited mitochondrial neurogastrointestinal encephalomyopathy, have been mapped to distinct nuclear regions (Nishino *et al.* 1999). Friedreich's ataxia and autosomal recessive hereditary spastic paraplegia (HSP) are good examples of diseases of secondary OXPHOS deficiency. HSP is characterized by progressive weakness and spasticity of the lower limbs due to degeneration of the corticospinal axons. HSP patients have two major morphological hallmarks of mitochondrial myopathy, RRF and COX-negative fibers (Casari *et al.* 1998). HSP is a genetically heterogeneous group of neurodegenerative disorders. An autosomal recessive form is associated with mutations in the gene encoding paraplegin, a novel mitochondrial protein highly homologous to the yeast mitochondrial ATPases AFG3, RCA1 and YME1, which have both proteolytic and chaperone-like activities at the inner mitochondrial membrane (Banfi *et al.* 1999). Friedreich's ataxia is an adolescent autosomal recessive disorder, the main clinical features of which include progressive ataxia, dysarthria, skeletal deformities, hyporeflexia, pyramidal features and hypertrophic cardiomyopathy (Schapira 1999a). Muscle biopsies from patients with Friedreich's ataxia also show typical mitochondrial changes. Genetic analysis has revealed an expanded intronic GAA repeat in the gene encoding frataxin (Koutnikova *et al.* 1997, Rötig *et al.* 1997). This frataxin deficiency results in mitochondrial iron accumulation, which probably results in oxidative stress and damage, which in turn may cause a decrease in mtDNA levels and severe deficiencies in the activities of complexes I–III (Schapira 1999b). Wilson's disease is an inherited disorder of copper homeostasis characterized by an abnormal accumulation of copper in several tissues, particularly the liver, brain and kidney. The disease-associated gene encodes a copper-transporting P-type ATPase located in the mitochondria (Lutsenko & Cooper 1998). Mitochondrial function has not been investigated in Wilson's disease but one would predict that an accumulation of copper may promote oxidative damage and result in an OXPHOS defect predominantly affecting complex IV (Lutsenko & Cooper 1998).

5.5. Disease associated with haplogroups

MtDNA is exposed to a higher mutation rate compared to nuclear DNA as it is not protected by histones. Furthermore, the repair mechanisms in mitochondria are not as efficient as those in the nucleus. Among the numerous mutations that have accumulated in mtDNA during evolution there are several ethnic specific single nucleotide polymorphisms (SNPs) that enable the definition of discrete and region specific subdivisions of the human population called mitochondrial haplogroups (mtHgs). Specific mtHgs might reflect functional differences in energy metabolism and were therefore linked with human diseases like neurodegenerative disorders or cancer. Polymorphisms in mtDNA might also represent modifier factors as has been shown in Leber's hereditary optic neuropathy (LHON) where mtHgs contribute to the phenotypic expression of mtDNA mutations (Carelli *et al.* 2006; Pello *et al.* 2008;Pierron *et al.* 2008)

5.6. List of mtDNA mutations and diseases

Disease	mtDNA mutation	Affected gene
LHON Primary mutations	G3460A	ND1
	G11778A	ND4
	T14484C	ND6
Intermediate mutations	G5244A	ND2
	G15257A	Citocrome b
Secondary mutations	T3394C	ND1
	T4160C	ND1
	T4216C	ND1
	A4917G	ND2
	G7444A	CO I
	T9101C	ATPasa 6
	G9438A	CO III
	G9804A	CO III
	G13708A	ND5
	G13730A	ND5
	G14459A	ND6
	G15812A	cyt b
NARP	T8993G	ATPasa 6
Leigh (MILS)	T8993G	ATPasa 6
	T8993C	ATPasa 6
MELAS	A3243G	tRNALeu(uur)
	C3256T	tRNALeu(uur)
	T3271C	tRNALeu(uur)
	T3291C	tRNALeu(uur)

	T9957C	COIII
MERRF	A8344G	tRNALys
	T8356C	tRNALys
Diabetes and deafness	A3243G	tRNALeu(uur)
Cardiomyopathy (MICM)	A3260G	tRNALeu(uur)
	C3303T	tRNALeu(uur)
	A4269G	tRNAIle
	A4300G	tRNAIle
	A4317G	tRNAIle
	C4320T	tRNAIle
	G8363A	tRNALys
	T9997C	tRNAGly
Mitochondrial myopathy	T3250C	tRNALeu(uur)
	A3302G	tRNALeu(uur)
	C15990T	tRNAPro
Deafness induced by aminoglycosides	A1555G	12S rRNA
Senso-neural deafness	T7445C	tRNASer(ucn)
Sideroblastic anemia	G12301A	tRNALeu(cun)
Symmetric multiple lipomatosis s	A8344G	tRNALy
CPEO	Unique deletion	
	A3243G	tRNALeu(uur
	A5692G	tRNAAsn
	G5703A	tRNAAsn
	C3256T	tRNALeu(uur
Exercise intolerance	G15084A	Citocromo b
	G15168A	Citocromo b
	G15723A	Citocromo b
	G14846A	Citocromo b
	Deletion of 24pb	Citocromo b
LIMM	A15923G	tRNAThr
Sudden death	A3251G	tRNALeu(uur)
Bilateral necrosis of striation	T9176C	ATPase 6
	T8851C	ATPase 6
Multi-systemic	A3251G	tRNALeu(uur)
	A3252G	tRNALeu(uur)
	C3256T	tRNALeu(uur)
Corea and dementia	G5549A	tRNATrp
LHON and dystonia	G14459A	ND6
Diabetes and myopathy	T14709C	ND6
Pearson	Unique deletion	
Kearns-Sayre's	Unique deletion	

Abelardo *et al.* (2001)

6. Mitochondrial DNA as a genetic marker

In humans, the mtDNA is a double-stranded, circular, covalent closed molecule of 16.5 kb. The mtDNA is inherited as a haploid from the mother and heteroplasmy has been found rarely. The human mitochondrial DNA (mtDNA) genome was sequenced and enumerated over twenty years ago (Anderson et al., 1981). The approximately 16,569 base pair genome encodes 13 polypeptides, 22 tRNAs, and 2 rRNA subunits. This closed, double-stranded, circular genome can be classified according to function: the coding region (about 15.5 kb of the genome) and the non-coding control region (about 1.1 kb of the genome). All of the genes in the coding region are highly concatenated and only occasionally have non-coding "spacer" sequences separating the genes. The control region has an important regulatory function for the mitochondria and contains sequences to initiate both transcription and DNA replication of the heavy strand. Many laboratories have focused on sequences within the non-coding control region of the mtDNA genome (mtGenome), specifically hypervariable regions I and II (HV1 and HV2), they span roughly positions 16024-16383 and 57-372, respectively (numbered correspondingly to the reference sequence, Anderson et al., 1981). Although the boundaries of HV1 and HV2 are not rigidly defined, the dense array of polymorphisms within HV1/HV2 has made this region an attractive target for sequencing studies of mtDNA variation. From a populational perspective, it could be considered as a system of small, sexually isolated demes, or clonal lineages, with an evolutionary rate 5 to 10 times faster than the nuclear genome. All these characteristics make this molecule ideal for evolutionary studies. Mitochondrial DNA sequence variation can be used to construct phylogenetic networks that enable more detailed analyses of evolution and population history. A phylogenetic network allows the relationships between individual sequences to be displayed and a time scale between sequences to be determined. The analysis of mtDNA has become an important tool for studying human population structure and history (Stoneking *et al.* 1993). Most phylogenetic analyses of mtDNA have been based on sequence variation in the first hypervariable segment of the control region (HVS-I), which is the most variable part of mtDNA. RFLP studies of the mtDNA coding region have been used to classify mtDNA haplogroups. Phylogenetic analyses that have been based on either RFLP data or

sequences from the HVS-I almost invariably fail to form nested sets of haplotypes, due to poor resolution (Bandelt *et al.* 1995). These RFLP data have subsequently been expanded, and the most accurate phylogenetic networks for European mtDNA have been constructed by using sequence data from HVS-I (Macaulay *et al.* 1999) or HVS-II sequences (Helgason *et al.* 2000) augmented with data from RFLP analyses of the coding region. Complete sequences, which have now begun to emerge, represent the best resolution that it is possible to obtain in phylogenetic analyses when analysing mtDNA evolution (Richards & Macaulay 2001). The analyses are complicated, however, by the presence of homoplasies, parallel mutation events or reversals. Homoplasies are substitutions occurring at an identical position but caused by separate mutational events. A network constructed by means of a median algorithm has been suggested as the best method to analyse mtDNAs. A median network generated using a table of binary data contains the same information as originally, but in a more illustrative way, so that all the data can be arranged in a novel network portraying mtDNA relationships. Labelled appropriately, the network can predict haplotypes, tell us where homoplasy is located, which sites are frequently subject to mutations, where a consensus sequence is and whether recombination is likely to have occurred. (Bandelt *et al.* 1995.) Since the median network harbours all the most parsimonious trees for the input data, it yields a concise picture of the data.

6.1. RFLP and Control Region Analyses

Early studies of mtDNA identified mitochondrial mutations based on Restriction Fragment Length Polymorphism (RFLP) data, pioneered by Wesley Brown and Doug Wallace (Denaro *et al.* 1981). The use of "low resolution" restriction analysis (using 5-6 restriction enzymes) revealed a phylogeny that was "star-like", with a central mitochondrial type shared among diverse populations throughout the world (Johnson *et al.* 1983). Using "high resolution" restriction analysis (12 restriction enzymes), an analysis by Cann *et al.* (1987) revealed a more highly resolved phylogenetic tree of mtDNA. In the late 1980s and early 1990s, as DNA sequencing became cheaper, easier, and faster to perform, mtDNA sequencing of the control region became an additional method utilized for genetic studies of human mtDNA evolution. The analysis of RFLP variation in global populations permitted the identification of a number of monophyletic

clades in which all mtDNAs could be classified. These clades, or haplogroups, were distinguished by ancient mutations that occurred thousands of years ago.

6.2. Molecular evolution studies using mtDNA

Analysis of mtDNA is commonly performed to investigate relationships between populations for determination of ancestry. The non-recombining property of mtDNA makes it a powerful tool for generating support of when and where the human species (*Homo sapiens*) originated. It has been shown by DNA analysis that the first modern humans evolved in Africa (mtDNA haplogroup L) 100,000 to 200,000 years ago. Furthermore, *Homo sapiens*have been estimated to spread from Africa, commonly described as the Out of Africa model, approximately 50,000 years ago (± 20,000 years). In addition, analysis of mitochondrial as well as nuclear DNA obtained from the extinct species Neanderthal, dated to approximately 38,000 years before present, provides aluable information concerning evolution and the origin of modern humans. The time since Neanderthals and humans diverged is estimated to be about 500,000 years ago. It is also suggested that Neanderthals did not contribute any mtDNA to present humans (Ingman *et al.* 2000; Green *et al.* 2006; Noonan *et al.* 2006).

6.3. Entire mtGenome sequencing

The emergence of "mitogenomics" began with the landmark publication of 53 complete human mtGenomes from diverse global populations in late 2000 by Ingman and colleagues. The robust phylogenetic tree from this global sample confirmed previous research (Cann *et al.*, 1987; Vigilant *et al.*, 1991) that humans migrated out of Africa to populate the world. Soon afterward, Finnila *et al.* (2000) used Conformation-Sensitive Gel Electrophoresis (CSGE) for analysis of coding region sequences of 22 Finnish individuals belonging to haplogroup U. Finnila *et al.* (2001) followed up this effort with a phylogenetic analysis of 192 mtGenomes from Finnish samples spanning all European Caucasian haplogroups. The maximal information afforded by complete mtGenomes has allowed a "coming of age" for mitochondrial gene trees (Richards and Macaulay, 2001).

As more laboratories generate entire mtGenomes, the fruits of mitogenomics are coming to bear in a number of areas. In addition to the confirmation of the out of African hypothesis (Ingman *et al.*, 2000 and Maca-Meyer *et al.*, 2001), the lack of recombination in mtDNA has been rigorously confirmed (Ingman *et al.* 2000) – despite a flurry of

debate to the contrary (Innan and Nordborg, 2002; Hagelberg, 2003). Entire mtGenomes have been used to discover novel polymorphisms associated with haplogroups (Herrnstadt *et al.*, 2002). Torroni *et al.* (2001) observed differences in the mutation rate of African haplogroup L2 subtypes. Mishmar *et al.* (2003) proposed that selection plays a role in the mtGenome variation among different haplogroups, and Meyer and von Haeseler (2003) have studied site-specific substitution rates in the entire mtGenome.

6.4. Mutation rate variation in the coding region-

6.4.1. Characterization of the relative mutation rates in HV1/HV2

One salient characteristic of human mtDNA evolution is extreme substitution rate heterogeneity from site to site (Wakeley, 1993). Several studies have sought to characterize the relative mutation rates of sites in the two HV regions (Wakeley, 1993; Malyarchuck *et al.*, 2002). The mutation rate spectrum in HV1 and HV2 is best described by a gamma distribution, with an "L-shaped" curve **(Figure 2).** In such a distribution, a few sites exhibit very high mutation rates (so called, "hotspots") while most sites have a range of low substitution rates (or are practically invariant). The shape parameter (α) of the gamma distribution is used as a measure of the amount of rate variation (reviewed in Yang 1996). The α parameter is inversely related to the extent of rate variation: low values of α (≤ 1) produce an "L-shaped" distribution while increased values of α (> 1) are indicative of intermediate rates (i.e. few sites are either invariant or fast). As α approaches infinity, all sites have the same rate Differential substitution rates have been shown to create artifacts in phylogenetic reconstruction (Tateno *et al.* 1994; Yang 1995) if not properly accounted for. Mutation rate heterogeneity can mimic population expansion (Aris-Brosou and Excoffier 1996) and bias estimations of population genetic parameters such as estimation of effective population size (Fu, 1994) and calibration of the molecular clock

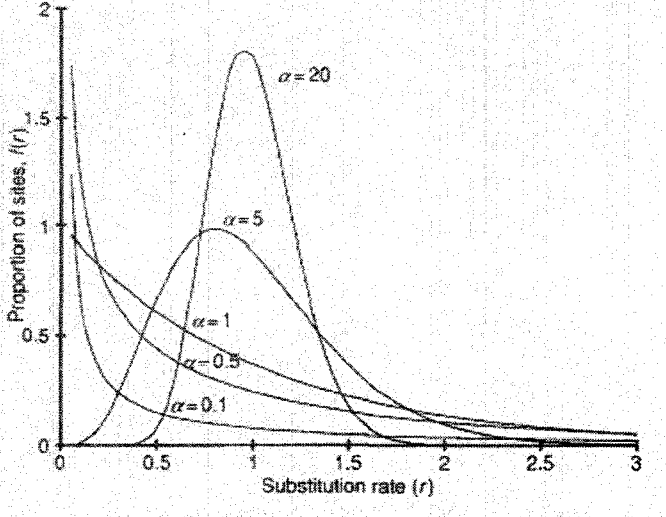


Figure 2. The Gamma Distribution of Substitution Rates.

(Yang, 1996). Wakeley (1994) demonstrated that ignoring the mutation rate variation could greatly underestimate the transition: transversion ratio.

The density function (proportion of sites) is graphed against the substitution rates at sites. The gamma distribution has a shape parameter, α, which is inversely related to the degree of rate variation. Low values of α (≤1) the gamma curve is "L-shaped", meaning that most sites are invariable, while a few sites have a fast mutation rate. As α becomes larger (>1), the curve becomes "bell-shaped", meaning that nearly all sites have an intermediate rate with a few sites invariant and a few sites fast. Figure adapted from Yang, 1996.

6.4.2. Methods to determine mutation rate variation

Several methods have been used to determine the mutation rate variation within the mtDNA control region. These techniques include accessing the relative mutation rates from phylogenetic trees evaluated using parsimony (Wakeley 1993; Hasegawa *et al.* 1993), maximum likelihood (Excoffier and Yang 1999; Meyer *et al.* 1999), and pair wise distance methods (Pesole and Saccone 2001). The determination of relative mutation rates using phylogenetic trees inferred by using parsimony has been shown to underestimate the true mutation rate by overestimating α (Yang, 1996). Maximum likelihood (ML) methods utilize an evolutionary model to estimate DNA substitutions and have been shown to perform better than parsimony-based methods to determine unbiased measures of α (Yang 1994). One disadvantage of the ML method is the intense computational requirements for analysis, limiting the number of sequences that can be evaluated at one time. A pair wise distance method has also been used to determine relative mutation rates (Van de Peer *et al.* 1993). This method is based on the idea that in comparing two sequences, the probability of observing a difference at a particular site depends on the site-specific mutation rate and the evolutionary distance between the two sequences. Thus, when comparing two very closely related sequences, there is a much higher probability that any observed nucleotide differences between the two sequences are a result of changes at fast evolving sites (Van de Peer et al., 1993; Meyer and von Haeseler, 2003). Since this method simply compares sequences and calculates the "disparity index" (Van de Peer *et al.*, 1993), a much larger number of sequences can be analyzed without having to rely on phylogenetic tree building methods (Van de Peer and De Wachter 1993; Pesole and Saccone 2001).

6.4.3. Assumptions of the rate variation in the coding Region

Characterization of the site-by-site spectrum of relative mutation rates in the coding region is now feasible with the recent, growing number of published mtGenome sequences. Until very recently, the assumption that the slower evolving coding region has a relatively homogeneous mutation rate appears to have been the rule. By assuming, "there is no evidence of variation in the mutation rate", this group concluded that the appearance of excessive homoplasy in the mtDNA coding region was due to recombination. Criticisms of the data used in that study (Macaulay et al. 1999b; Kivisild and Villems 2000), the methods used (Jorde and Bamshad 2000; Kumar *et al.* 2000), and recent analyses of entire mtGenomes (Ingman et al. 2000; Elson *et al.* 2001; and Herrnstadt *et al.* 2002) have abrogated the controversy of recombination in mtGenome (Innan and Nordborg 2002). There appears to be inconsistency in the literature as to the relative amount of homoplasy in the coding region. For example, Herrnstadt *et al.* (2002) notes the "relatively large number of sites" showing the "extensive" amount of homoplasy in their data set of 560 coding region mtGenomes. On the other hand, another group (Yao *et al.* 2003) examines the same data and notes that, "Homoplasy in the coding region is much less than in the control region and may have only a few hot spots." Meyer and von Haeseler (2003), using the pair wise distance comparison method between pairs of sequences (Van de Peer *et al.* 1993), identified site-specific mutation rates and estimated the parameter for the mtGenome. In the pair wise method, Van de Peer et al. (1993) first estimated genetic distances between pairs of sequences to identify sites that have undergone multiple mutations. Then, Van de Peer *et al.* (1993) used probabilistic analysis to determine the site-specific rates at each position. Meyer and von Haeseler (2003) determined that the statistical basis used by Van de Peer *et al.* (1993) was not well established. Meyer and von Haeseler (2003) used the pair wise distance method as in Van de Peer *et al.* (1993) followed by a maximum likelihood model to estimate the site-specific rates in the mtGenome.

7. Phylogenetic tree building methods

Different tree-building methods exist for the study of phylogenetic relatedness between DNA samples (Li 1997). Most generally the algorithms can be divided into groups: those relying either on (i) distance, e.g. neighbor joining (NJ), or on (ii) character state, e.g.

maximum parsimony (MP) and maximum likelihood (ML), analyses.

7.1. Neighbor joining method

NJ tree construction (Saitou and Nei 1987) starts with a starlike tree and a distance matrix inferred from pairwise comparison of all sequences. Sequentially, the search for the closest neighbors in the matrix defines additional interior branches in the tree, minimizing its total length.

7.2. Parsimony method

MP methods (Fitch 1977; Swofford 1993) search for tree(s) that require the smallest number of substitutions. Tree construction employs only informative sites - substitutions that favor certain branching order; and from the set of all possible or heuristically chosen samples of trees the one(s) with the smallest number of substitutions (character state changes) are chosen.

7.3. Maximum likelihood method

ML methods (Felsenstein 1988) assign likelihood values to trees or subsets of a tree to choose the best tree by the value. The likelihood estimates are derived from a probabilistic model that is specified for character state changes. Such models take into account substitution rate from one character state to another. The rates can be taken as uniform for all substitution types (Jukes and Cantor's 1-parameter model), or different values can be given for transitions and transversions (Kimura's 2-parameter model). Rates can be further refined by different substitution types and GC content. Unlike MP method ML analysis makes use of all sites available in the sequences.

8. Phylogenetic networks

Multiple hits are a common problem in all phylogenetic analyses as they cause incompatible character states. The higher the number of pairwise incompatible (homoplasious) sites the higher is the number of trees with equal length that can be drawn from the data set. One particular tree from such a forest of MP trees alone, thus, can be misleading as far as character conflicts are resolved arbitrarily. At the cost of lower resolution the given homoplasies can be expressed by alternative pathways in a form of networks. Fast algorithms of network construction have been developed, relying either on sequential split decomposition of each informative character in the sequence matrix (Bandelt 1994; Bandelt *et al.* 1995), ending up with a (reduced) median network

containing (almost) all MP trees, or on sequential introduction of inner branches between components of shortly connected nodes (Bandelt *et al.* 1999).

9. Rooting of phylogenetic trees

Any tree construction produces unrooted trees which do not define directions of evolutionary relationships nor make any assumptions about the common ancestor of the group. The tree can be rooted if additional knowledge is supplied from an outgroup. The latter is usually chosen from another group that is externally known to have separated from the common phylogenetic lineage before the existence of the most recent common ancestor of the group under study. If such information is not available, an alternative, yet not so powerful, way of rooting a tree is to specify a node in the tree from which the distance to all terminal nodes is minimal as the midpoint root. Insertion of parts of mtDNA into the more conservative nuclear genome can serve as a special case of outgroup - a recent insertion of a D-loop segment into chromosome 11 has been proved very useful for rooting human mtDNA tree (Zischler *et al.* 1995).

10. Mutation rate calibration

One of the crucial points in mtDNA based population genetics is the calibration of the molecular clock. Since timing is decisive for the interpretation of the demographic history of populations, a reliable relation between sequence diversity and the time scale is needed. Several approaches have been taken, all based on assumptions that can be quantitatively checked, like (i) constant rate in different lineages, (ii) neutrality of the mutations being used. Two of them rely on 'outside' information, like fossils or historical records, which have to be taken as they are.

1. One way of clock calibration is to consider variation in populations from specific geographical regions with a well-known time of colonization.The extent of differentiation within clusters specific to New Guinea, Australia and the New World have been used to estimate the mean rate of mtDNA divergence within humans. It has been found that the divergence rate (twice the substitution rate), averaged over the whole mtDNA molecule, lies between 2-4% (Cann *et al.* 1987; Torroni *et al.* 1994c; Wilson *et al.* 1985) and for transitions in a HVS-I segment (16,090-16365) is about 36% per million years (Forster *et al.* 1996).

2. An out-group or interspecies calibration method considers differentiation within one

species compared to the distance from the MRCA with another related species. The date for the split of these species is taken from paleontological evidence and related with the sequence diversity. Horai (Horai *et al.* 1995), for example, considered the divergence time of African apes from fossil record as being 13 million years and estimated from that figure that the human/chimpanzee split occurred 4.9 million years ago. This calibration gave the substitution rate at synonymous sites and in the D-loop region as 3.89 and 7.00 x 10-8/site/year, respectively. From these estimates the age of the last common ancestor of the human mtDNAs was inferred as 143,000 ± 18,000 years. Similarly, the distance between humans and chimpanzees was used to calibrate the rate of most widely used 360 bp segment of HVS-I (Ward *et al.* 1991), yielding the divergence rate of 33% per million years. For the whole control region, with a total of 751 nps, 23% per million years of divergence has been estimated (Stoneking *et al.* 1992).

3. Pedigree calibrations try to estimate the mutation rate directly from samples of known genealogy. Fast rates (260% divergence per million years) obtained from the first few pedigree studies (Parsons *et al.* 1997) were taken as being in conflict with those calculated phylogenetically. It was argued (Paabo 1996) that the pedigree rate could be more appropriate to use in studies focused on events in a timescale of hundreds or thousands of years. Now, when more data have became available (Bendall *et al.* 1996; Jazin *et al.* 1998; Soodyall *et al.* 1997) the pedigree estimates have become close to the conventional ones and an improved understanding of the differences between the mutation and fixation rates, and differences in rates between sites in control region, suggests that there is no need to revise the mutation rate estimates (Macaulay *et al.* 1997).

11. Calibration of the mtDNA molecular clock

The molecular clock hypothesis holds that for any given DNA sequence, the rate of evolution is approximately constant over all evolutionary lineages. This regularity of molecular evolution would stand in direct contrast to the non-uniform change of morphological evolution. In principle, this hypothesis could be tested using known calibration points of dated lineage divergences. This requires accurate, independent dating of the lineages being investigated by other disciplines, most notably palaeontology. Mutation rate calibration can either be performed directly on individual meioses, or indirectly through the observation of a certain amount of divergence across a known time

span. The latter approach often uses divergence between species whose divergence is well dated in the fossil records.

One of the approaches is to estimate the mean rate of mtDNA divergence in different geographical regions for which approximate time of colonization is available from historical sources. The extent of differentiation within clusters specific to New Guinea, Australia and the Americas has been used and the divergence time (twice the substitution rate) estimate between 2-4% per million years has been calculated for whole human mtDNA molecule (Cann *et al.*, 1987; Torroni *et al.*, 1994a). The same value fro transitions of D-loop HVS-I region (between nps16090 to 16365) was found to be 36% per million years (Forster *et al.*, 1996).

The other opportunity is to use the outgroup method, which compares the average amount of sequence variation between two species considering their distance from the MRCA. The time of their split is taken from paleontological evidences. Recently, Kivisild *et al.*, (2006) have performed calibration using nuclear inserts of mtDNA and the chimpanzee consensus sequence as outgroups appeared between haplogroup L0 and the rest of the phylogeny. Extensive interspecies homoplasy and mutational saturation was highlighted by the fact that for more than one-third (417/1292) of the variable sites, regardless of their phylogenetic position on the tree, the derived allele among humans corresponded to the chimpanzee allele. In agreement with noncoding region information (Aquadro and Greenberg, 1983), a high ratio (21.5 on average, 34.8 in synonymous positions) of transitions to transversions was observed in the coding region (577–16023). Interspecies calibration of the molecular clock over the complete mtDNA sequence (Ingman *et al.*, 2000; Mishmar *et al.*, 2003) is problematic because of saturation of transitions at silent positions and the effect of selection on the fixation rate of amino acid replacement mutations . Assuming 6 million years for the human–chimpanzee species split (Goodman *et al.*, 1998) and 6.5 million years for the most recent common ancestor of their mtDNA lineages (Mishmar *et al.*, 2003), Kivisild *et al.*, (2006) estimated the average transversion rate at synonymous and rRNA positions as 2.1×10^{-9} and 4.1×10^{-10} / year/position, respectively. Using the observed relative rates of different substitution types in humans, the average transition rate at 4212 synonymous positions is 3.5×10^{-8} (SD

$0.1x10^{-8}$)/year/position. Over all genes in mtDNA this would be equivalent to accumulation of one synonymous transition/6764 (SD 140) years on average. The coalescent date of the human mitochondrial DNA tree using this rate is 160,000 (SD 22,000) years. This coalescent date is broadly consistent with the dates of the *Homo sapiens* fossils recognized so far from Ethiopia (Clark *et al.*, 2003; White *et al.*, 2003; McDougall *et al.*, 2005). The TMRCA of all the Eurasian, American, Australian, Papua New Guinean, and African lineages in clade L3 dates to 65,000 ± 8000 years while the average coalescent time of the three basic non-African founding haplogroups M, N, and R is 45,000 years (Kivisild *et al.*, 2006). These estimates, bracketing the time period for the recent out-of-Africa migration (Stringer and Andrews, 1988), are younger than those based on calibrations involving all coding region sites (Ingman *et al.*, 2000; Mishmar *et al.*, 2003) but are still in agreement with the earliest archaeological signs of anatomically modern humans outside Africa (Mellars, 2004). The differences between the date estimates of previous studies are most likely due to the overrepresentation of possibly slightly deleterious nonsynonymous mutations in the younger branches of the tree (Elson *et al.*, 2004) that introduces a bias to the coalescent approach if all the sites of the coding region are used.

12. Pitfall of Age estimation

For a single locus, there has been a quite spectacular breadth of age estimates for the TMRCA of mtDNA. In some cases, the explanation for these divergent estimates is easy to spot. There has been an almost perennial confusion between *substitution rates* and *divergence rates* that seems to remain no matter how many times it is pointed out. The former measures the rate of change from an ancestral sequence to a contemporary descendant sequence, whereas the latter measures the rate of change between two sequences that diverged from a common ancestor. So, reading the substitution rate as a divergence rate would slow down the evolution by a factor of a half, whereas the other way round the evolutionary rate would get doubled. This can cause particular confusion when it arises in a work that aims to compare different estimates of the mutation rate in the control region (Santos *et al.* 2005). Perhaps the overestimates have caused the greatest difficulties. The brave multiregionalist, confronted with the shallow coalescence time usually estimated for human mtDNA, is easily tempted to refer to some of the more

extreme age estimates yielding up to half a million years or so, obtained in the mid-1990s and based on meagre data and dubious methods. Such outlier estimates are then labelled 'conservative' and are still being woven into multiregional models to this day (Eswaran *et al.* 2005). There are, however, more recent and younger age estimates which equally lack any solid basis.

13. Weekness of mtDNA analysis

Whilst the knowledge that can be gained from mtDNA analyses is significant, the possible drawbacks associated with basing scientific conclusion on the analysis of such a genetic marker must not be overlooked. mtDNA represents only one locus within the genome (Ingman *et al* 2000), and provides information relating to the maternal line of descent only (Salas *et al* 2005). In addition, interpretation of mtDNA is 'notoriously difficult' (Bandelt *et al* 1995), not least due to the abundance of homoplasy within the data (Brandstätter *et al* 2004); also known as parallel mutation or reversal, this is rife throughout mtDNA sequences, especially within the control region (Eyre-Walker & Awadalla 2001), with Howell and colleagues (2004) describing the high number of control region sites displaying apparent homoplasy as 'the most striking result' of their study. Homoplasy results either from recombination, or via the same site acquiring the same mutation in different parts of the tree (Eyre-Walker & Awadalla 2001); with the occurrence of the former already shown unlikely, it has been argued that the high rates of homoplasy within mtDNA provide evidence for significant variation in the mutation rate within the mitochondrial genome (Eyre-Walker & Awadalla 2001, and references therein). In addition, when focussed specifically on human origins or the earliest of human dispersals, these analyses are further complicated by the greater time-depth of African lineages, compared to that of others around the globe (Salas *et al* 2002); modern humans are believed to have evolved in Africa around 200,000 years ago (Cann *et al* 1987), before radiating out and replacing archaic human forms in Europe and Asia, with the L3 haplogroup giving rise to the widespread M and N mtDNA lineages (Macaulay *et al* 2005). Despite the frequency with which intraspecific mtDNA variation analysis is now used as an investigatory tool, the interpretation of such data is far from error-free. Often as a result of homoplasy, reticulations can also frequently find their way into networks via reference base error and base-shift error (Bandelt *et al* 1995). In some cases, errors

such as these can be classed as 'phantom mutations' – unfortunately highly widespread mutations generated via the sequencing process, frequently due to misreading (Bandelt *et al* 2002) or post-transcriptional error (Brandstätter *et al* 2004). Such errors can have a significant detrimental effect on the accuracy of any results gained from data under consideration.

14. Recombination within mtDNA genome?

The mtDNA genome is frequently reported as a non-recombining genetic marker (Herrnstadt *et al* 2002). However, as our knowledge of the mtDNA region has grown, this understanding has been questioned. Eyre-Walker and Awadella (2001) propose three possible routes via which recombination within the mtDNA genome could take place; each case challenges the consensus that as mtDNA is inherited maternally, it is therefore clonal. Of the three pathways – recombination via heteroplasmy, nuclear pseudogenes and paternal leakage – the authors conclude the latter to be the most likely, stating that whilst evidence is limited, enzymes required for homologous recombination are encoded within the mitochondria, and even extremely sporadic recombinational events would have a 'significant and detectable effect in terms of evolution'. However, the authors have been criticised for the methodologies employed, with limited samples in three of their data sets (Jorde & Bamshad 2000), and a measure ($r2$) prone to producing misleading results (Kumar *et al* 2000). Overall, a general consensus would view mtDNA recombination as non-existent (Ingman *et al* 2000).

15. Homoplaseis and multiple hits

In 1983 Aquadro and Greenberg (Aquadro and Greenberg 1983) found that the sequence diversity in the D-loop region of mtDNA is several fold higher than in the coding parts and that there exists a 32-fold bias of transitions over transversions. Wherefore transitions, in a relatively long run, tend to get 'saturated' and are about to undergo multiple hits, which means that one site may have gone through many substitutions and yet be at the same state. Multiple hit corrections therefore are to be made whenever attempting to draw phylogenetically reasonable conclusions, especially those concerning time estimates. One way of solving this problem for transitions is to compare the observed difference between two species with the one expected from transversional distance under a model of known ratio of transitions and transversions (Vigilant *et al*. 1991). Another way to cope with the

high levels of homoplasy in HVS-I haplotypes, caused by either 'visible' parallelisms or 'hidden' multiple hits, is a combined application of both RFLP and HVS-I data (Starikovskaya *et al.* 1998; Torroni *et al.* 1996).

16. Rate variation

Besides 20-30 fold transitional bias it has been noted that the rate variation between sites from the control region is also significantly high (Ohno *et al.* 1991; Wakeley 1993). Transitions at sites like 16093, 16129, 16209, 16311 and 16362 from HVS-I and 00146, 00150, 00152 and 00195 from HVS-II occur in many different lineages and these sites can be considered as mutational hotspots. The heterogeneity is higher in HVS-II (Aris-Brosou and Excoffier 1996) that contains a handful of extremely fast evolving sites and a long list of those that rarely undergo any mutation. Rate variation can be taken into account in phylogenetic homoplasy solving by giving different weights to sites according to known rate variation (Helgason *et al.* 2000).

17. Mismatch distributions and demographic expansions

Pairwise comparison of sequence types gives a frequency distribution of distances ("mismatch distributions", Harpending *et al.* 1993) between all possible pairs of sequences in the sample. When a population goes through a rapid expansion then most of the variation within that population is preserved and new mutations get fixed with higher probability. Given the random nature of mutation cumulation, the frequency distribution of pairwise distances is thus expected to be unimodal and fit the Poisson process. Contrarily, in populations with constant size the distribution is expected to be a multi-modal or "bumpy" one. Thus, in the ideal case there is a downright correlation between the population demographic history and the mismatch distributions. In reality, very often, the population history can be a sum of several fusions, splits and multiple expansions or bottle-necks. Hence, a simplistic analysis of mismatch distributions in a whole sample can be misleading. For example, African populations were first interpreted (Graven *et al.* 1995; Watson *et al.* 1996) as stationary because of their bumpy mismatch distributions. Yet, when phylogenetically properly dissected (Watson *et al.* 1997) several African-specific lineages clusters were detected that exhibited unimodal mismatch frequency pattern reflecting possible ancient expansions. These separate clusters had expanded and were further fused together before the emergence of modern ethnic

groups.There is a contributive correlation between the coalescence age estimate of a tree and the effective population size for a given population since the coalescence. By theory, the smaller the effective population size is the sooner a relevant tree is expected to coalesce in the past. For example, human mtDNA diversity compatible with the African "Eve's" age of 200,000 years has been argued to infer that the minimal effective number of females in the late Middle Pleistocene could have been in the order of only one to ten thousands (Harpending *et al.* 1993). Further, the peaks of mismatch distributions suggest that a major population expansion in human history took place long after that time, around 60,000 years ago (Harpending *et al.* 1993; Mountain *et al.* 1995; Rogers and Harpending 1992; Rogers and Jorde 1995; Sherry *et al.* 1994).

18. Early studies of phylogeographic diversity of human mtDNA

The first application of mtDNA to elucidate the origin of modern humans took place in the late 1970s and early 1980s, when Brown *et al.*, (1980) discovered the restriction fragment pattern of mtDNAs among individuals from diverse ethnic and geographical origin differed substantially. On the basis of an estimated rate for base substitution of 1%/106 year (Brown *et al.*, 1979), man could have speciated or passed through a severe population constriction as recently as 180,000 years ago. The complete sequence of mitochondrial genome was published in 1981 (Anderson *et al.*, 1981) and soon after that many studies of different population were undertaken. While Brown *et al.*, (1980) had used 18 restriction endonucleases and 21 samples. Denaro *et al.* (1981) treated a large number of samples (235) from five ethnic groups with only a single restriction enzyme. They found that the presence of *Hpa*I restriction site that corresponds to the transition at np 3954 separates most of the Africans from the Eurasian individuals. Notably, the Africans had a derived sate of this np when compared to other primates, whereas the rest of studied population shared an ancestral state. Based on these findings was proposed, Asia is possible genetically central to the mtDNA radiations that are thought to have given rise to the human ethnic groups. This study was supported by the investigation of Blanc *et al.*, (1983) who discovered the frequent presence of ancestral state of one *Hinc*II polymorphism (corresponding to transition in np 12406) among Asians and was upheld by studies of Nepalese (Brega *et al.*, 1986) and Chinese (Yu *et al.*, 1988) populations. Phylogenetic analysis of the Japanese population concealed a considerably high degree of mtDNA diversity (Horai *et al.*, 1984). Cann *et al.*, (1982) demonstrated that mtDNA diversity among aboriginal Australians is as diverse as in

any other populations tested in the Old World. Johnson *et al.*, (1983) showed that the mtDNA diversity is greatest in Africa, and that all mtDNA variants present toady can be seen as deriving from a single phylogenetic tree. Although they argued that the highest diversity in Africa can be caused by the longer age of African variants as well as by different mutation rates in different mtDNA lineages, the mid-point root of their mtDNA tree indicated the African origin. Later, Excoffier and Langaney (1989) suggested that the high diversity of African could be explained equally well by assuming that selection has played a significant role in creating the present mtDNA variation. The possibility to trace the origin and to study the demographic history of human populations by the use of mtDNA as a population specific marker became highly visible in 1987 (Cann *et al.*, 1987). They used high-resolution restriction analysis of 147 mtDNAs from five geographic populations and concluded that the root of human mtDNA phylogenetic tree is in Africa as well as all the populations examined except the African populations have multiple origins, implying that each area was colonised repeatedly. Compared to earlier studies (cited above) that allowed the screening of 2-4% of the total mtDNA sequence variation, in this analysis a ten time higher resolution level was obtained.

18.1. Multiregional model of human evolution

Nearly at the same time when Chris Stringer and Alan Wilson molded the theory of recent African origins, Milford Wolpoff, Alan Thorne and Wu Xinzhi put forward another explanation for the pattern of human evolution (Thorne and Wolpoff 1992). The multiregional model of evolution traces all human populations back to a *Homo erectus* ancestor that supposedly left Africa at least 1 million years ago. Further, despite an everlasting network of intercontinental gene flow, each continental group of populations developed to *sapiens* in its own way, retaining in the process distinguishable features (and genes) of the particular local pre-*sapiens* populations. By the use of different analysis techniques and extended mtDNA data sets the theory of recent African origins of human mtDNA phylogeny has found overwhelming support (Penny *et al.* 1995; Watson *et al.* 1997). Moreover, most nuclear data sets also back up the theory (Nei and Takezaki 1996; Tishkoff *et al.* 1996) as do the paleoanthropological findings (Foley 1998). The global data base of mitochondrial DNA variation allows now to face in more detail the questions related to the MRCA, her age and the time and route(s) of expansions and dispersals in and out of Africa.

18.2. Recent African origin of " Mitochondrial eve" and the anatomically modern Humans

An opposite model to "multiregionalism" is the "recent out-of-Africa" model, known also as the *Garden of Eden* model, according to which AMHs evolved only in Africa and replaced all other preexisting archaic hominids. This theory was proposed in its extreme form — assuming total replacement of all archaic ancestors — by archaeologists Stringer and Andrews (1988) at about the same time when the genetic evidences of recent African origin of human mtDNA appeared (Cann *et al.* 1987). As for today, the recent out-of-Africa model has found preponderant support from the mtDNA analysis of large datasets analyzed with different methods (Horai *et al.* 1995; Ingman *et al.* 2000) as well as from the studies of Y-chromosome (Hammer 1995; Underhill *et al.* 1997; 2000), autosomal loci (Armour *et al.* 1996; Nei and Takezaki 1996; Jorde *et al.* 2000) and from paleoanthropological findings (Stringer and Andrews 1988; Foley 1998; Stringer 2000; 2003). However, this model is certainly over-simplified, as it does not account for continental subdivisions, environmental changes affecting the demographic history of populations, potential sex-linked differences in generation time *etc.*, and lacks a mechanism to explain the present diversity of AMH (Excoffier 2002). Phylogenetic studies on mtDNA in Europeans (Torroni *et al.* 1996; Macaulay *et al.* 1999*b*), Asians (Torroni *et al.*1993*a*; 1994*b*; Schurr *et al.* 1999; Kivisild *et al.* 2002; Yao *et al.* 2002*a*; 2002*b*; Kong *et al.* 2003*a*; 2003*b*), and Papuans (Stoneking *et al.* 1990; Forster *et al.* 2001) and Native Americans (Torroni *et al.* 1993*b*; 1994*a*; 1994*c*) have confirmed that each continent has distinct sets of phylogenetically deep mtDNA branches that speak in favour of the so-called *weak Garden of Eden* model (Harpending *et al.*1993). According to this model, present genetic variation has not been the result of uninterrupted demographic expansion of the out-of-Africa founders but rather by the different maturation and range expansion phases of regional gene pools that followed the initial expansion (Forster *et al.* 2001; 2004). Various alternative modes and pathways of dispersal of AMHs, following the northern (over Sinai) and/or southern (over southern Arabia) routes have been put forward, based on paleoanthropological (Lahr and Foley 1994; 1998; Stringer 2003) and genetic data (Cavalli-Sforza *et al.* 1994; Hammer *et al.* 1997; 1998; Jin *et al.* 1999; Kivisild *et al.* 1999*a*; 2000; 2003*a*; Quintana-Murci *et al.*

1999; Underhill *et al.* 2000; Cann 2001; Templeton 2002). Thus, modern humans originated from a recent single evolutionary event, whereas modern human diversity is the result of multiple evolutionary events brought about by multiple geographic dispersals.

19. The pioneer settlement of Modern humans in Asia

In order to dig more deeply into discussions about pioneer settlement of Asia, it is necessary to emphasize here that many recent genetic, archaeological, and anthropological studies have started to favour the Southern Coastal Route (SCR) concept as the main mechanism of the primary settlement of Asia (Lahr and Foley 1994; Quintana-Murci *et al.* 1999; Stringer 2000; Kivisild *et al.* 2003, 2004; Oppenheimer 2003).

19.1. Introduction

The coastal habitat as the medium for humans to penetrate from East Africa to Asia and Australasia was perhaps first envisaged by the evolutionary geographer Carl Sauer, who considered the populations taking this route as adapted to the ecological niche of the seashore (Sauer 1962). After reaching Southwest Asia, modern humans had a choice of two potential routes by which to colonize the rest of Asia. These two were separated by the world's mightiest mountain system—the Himalayas. The pioneer settlers could continue taking the SCR or they could change their habitat and turn instead to the north, passing through Central Asia and southern Siberia (or via the route that later became known as the Silk Road). Here, one has to avoid confusion with the 'Northern Route' of the out-of-Africa exodus and use the term 'Northern Asian Route'.

19.2. Palaeclimatologial context

The patterns of colonization of Asia by anatomically modern human (AMH) population(s) were undoubtedly highly dependent on the surrounding environment. Our knowledge of past climates is frozen in ice sheets: polar ice cores and (ocean) sediments have been intensively studied in order to reconstruct the climate of the past. These global changes are reflected in the more detailed regional palaeovegetation surveys based, for example, on ancient pollen analysis (Adams and Faure 1997; Adams *et al.* 1999; Ray and Adams 2001). After the Eemian interglacial, some 110 000–130 000 years ago, the global climate cooled until the period of the lowest temperatures, during the Last Glacial Maximum

(LGM) 15 000–25 000 years ago. This process was not a steady one; instead, there were multiple oscillations of warmer and chillier periods. Intense cold and arid, but short-lived, Heinrich-type events characterized the otherwise gradually cooling phase between 110 000 and 70 000 years ago. This was followed by the stage 4 Glacial Maximum (also known as the Early Wisconsin Glacial) extending to about 50 000 years ago with conditions rather similar to the LGM (Adams *et al.* 1999). Warmer but highly variable temperatures were characteristic of the period thereafter, extending until the onset of the LGM. As much as the changing temperatures, the peopling of Asia by AMHs was affected by the accompanying fluctuations in humidity. Lower temperatures generally mean less evaporation. The resulting global decrease in rainfall contributed to the extension of desert areas, for example, in Central and southwestern Asia. Even around 50 000 years ago, when a warmer and moister stage opened the green passage between the Arabian Sea and the Levant (the Zagros corridor), the deserts in Central Asia and northern Africa remained difficult habitats for most of the creatures—including humans (Oppenheimer 2003).

19.3. Archaelogical and palaeontological evidence of the peopling of Asia by AMH

Any fossil record is, inevitably, incomplete. Fossilization of skeletal remains is a rare event, depending on climate and probably many other factors, whereas their recovery depends on the intensity of archaeological investigation of a region. There is a particular issue when one considers the course of the likely coastal route out of Africa. If the beachcombing modern humans, being dependent on a seashore environment, indeed began colonizing Eurasia via the SCR, then many of the potential archaeological sites are at present submerged under the sea. An 80-m rise of the sea level (the difference between the sea level 70 000 years ago and today) altered the coastline considerably, shifting it hundreds of kilometres inland and probably inundating the range of beachcombing AMHs. Another important factor that needs to be considered is the tectonics of the continental shelves. Furthermore, the accuracy of fossil dating techniques is in constant dispute (Chen and Zhang 1991; Klein 1999). Despite these problems, fossils are and will probably continue to be the best evidence of the spread of AMHs around the globe.

19.4. How to reconstruct "pioneer settlement" from extant mtDNA diversity

An obvious starting point for deducing the patterns of the pioneer human settlement from

the extant mtDNA diversity is to identify regionally autochthonous haplogroups and calculate their coalescence ages. The average over the oldest of these would indicate the lower bounds for the start of the colonization. A founder type is identified as an ancestral node which is present (or may have been lost but is then phylogenetically reconstructed) both in the source and in the destination area (Richards *et al.* 2000). Ideally, the coalescence time of the founder type in the destination area would suggest the time of its arrival (Stoneking *et al.* 1990; Torroni *et al.* 1993a, b; Sykes *et al.* 1995; Forster *et al.* 1996; Richards *et al.* 2000).

19.5. The peopling of Asia as seen through the lens of mtDNA diversity

Complete and partial mtDNA coding-region sequences have been used to map the backbone and determine the fine structure of the mtDNA lineages present in Asia (Kivisild *et al.* 2002; Yao et al. 2002; Kong *et al.* 2003; Metspalu *et al.* 2004; Palanichamy *et al.* 2004; Quintana-Murci *et al.* 2004). The recent analysis of complete mtDNA sequences from 672 Japanese individuals has provided a significant refinement of the East Asian mtDNA phylogeny (Tanaka *et al.* 2004). The macrohaplogroups M and N effectively cover the whole mtDNA pool in Asia. The start of their dispersal has been dated to approximately 60 000–65 000 years ago (Maca-Meyer *et al.* 2001; Mishmar *et al.* 2003; Palanichamy *et al.* 2004). MacrohaplogroupM is slightly more frequent than N in Siberia, northern China, Japan, and South Asia, while in Southeast Asia it is the other way around. M is nearly absent from Southwest Asia, where subhaplogroups branching from the N (including R) trunk dominate the mtDNA landscape. The N and R subbranches in West and East Eurasia do not overlap, and they formtwo distinct mtDNA 'domains'. With approximately similar shares, these two make up most of the mtDNA pool of Central Asia (Comas *et al.* 2004). While the mtDNA makeup of the Americas represents an offshoot of the East Asian domain (Torroni *et al.* 1993a, b, 1994; Forster *et al.* 1996), Sahul (Australia/New Guinea) largely constitutes yet another autochthonous one. Stemming from the trunks R and M, haplogroups P and Q, respectively, cover more than half of the extant mtDNA pool sampled in Papua New Guinea (Forster *et al.* 2001). From the published full sequences (Ingman and Gyllensten 2003), additional Sahul-specific basal branches of M and N—namely S and O, which were baptized in Palanichamy *et al.* (2004)—are confirmed. Thus, both basal trunks of the Eurasian mtDNA tree show

deep-rooting Sahul-specific branches. South Asia, with its own specific branches of M and N, represents the third mtDNA domain in Asia. Haplogroups M2 and R5 and sub groups U2a, U2b, and U2c of U2 (Kivisild *et al.* 1999a, 2003; Quintana-Murci *et al.* 2004), which make up more than 15% of the South Asian mtDNAs, each show coalescence times of over 50 000 years (Metspalu *et al.* 2004). These haplogroups form a set of the most ancient Indian-specific haplogroups identified so far. A number of novel Indian-specific basal N and R lineages (N5, R7, R8, R30, and R31) were recently identified from complete sequences (Palanichamy *et al.* 2004). The phylogeography of these in South Asia needs further attention, but, significantly, their autochthonous presence in India clearly demonstrates that all the basal trunks—M, N, and R—have diversified in situ. The coding-region-based downstream classification of haplogroup M lineages in South Asia is on the way (Sun *et al.* 2006), testifying to considerable basal diversity and confirming that the M subhaplogroups of South Asia are different from those of East Asia (Kivisild *et al.* 1999b; Metspalu *et al.* 2004). Overall, then, the South Asian mtDNA pool consists of autochthonous branches of the global mtDNA tree that stem directly from each of the basal trunks M, N, and R. Note that the only major Indian-specific lineages not stemming directly from the trunk are the Indian subhaplogroups of U2, which may have a sister group U2e in West Eurasia (although we note that this putative sister relationship hinges upon a single transition, at nucleotide position 16051 in the control region, which may not have been a unique event at the base of haplogroup U). The divergence time of these U2 daughters reaches 50 000 years (Kivisild *et al.* 1999a). Meanwhile, haplogroups R2, U7, and W represent an intriguing link between the West and South Asian mtDNA pools. Their spread and coalescence times suggest pre-LGM gene flows in the area spanning from western India and Pakistan up to the Near East. As judged from the coalescence times of the region-specific subclades of these haplogroups, this genetic continuum was apparently interrupted by the expanding deserts in eastern West Asia during the LGM (Metspalu *et al.* 2004). Like Sahul and South Asia, the East Asian mtDNA pool is made up of autochthonous offshoots of M and N, most of which show coalescence times exceeding 50 000 years (Kivisild *et al.* 2002; Yao *et al.* 2002). While in South Asia we see a number of basal haplogroups branching from trunk R, in East Asia, only a few haplogroups, B (plus R11) and R9

(including F), spring out from the founder haplotype of haplogroup R . The putative monophyly of a supergroup R11_B is based solely on transitions at unstable nucleotide positions 16189 and 16519, which could very well have happened in parallel. Furthermore, no more than two East Asian haplogroups, A and N9 (including Y), trace back to MRCAs in the N trunk. As in South Asia, the richest trunk in East Asia in terms of haplogroups stemming from it is M. Complete mtDNA sequencing has indicated that Southeast Asia also harbours some autochthonousM, N, and R lineages apparently not found further north, in East Asia (Macaulay *et al.* 2005; Merriwether *et al.* 2005).

19.6. Conclusion

It appears that Central Asia and southern Siberia were not involved in the initial peopling of the continent. It is also evident that the initial fast train phase was followed by a long-lasting freezing of the major geographic pools of maternal lineages in the south and further gene flows northwards from Southeast Asia and subsequently back westwards along the Steppe Belt extending from Manchuria to Europe. At present, western Siberia, the Urals, and Central Asia form a huge continuous admixture zone encompassing East and West Eurasian maternal lineages—a process that has so far had only a minimal influence on the essentially distinct autochthonous patterns of mtDNA variation in most of South Asia, East Asia, Southeast Asia, and Australasia.

20. Global mtDNA variation, nomenclature of mtDNA haplogroups

The first studies of mtDNA variation in human populations used either the RFLP analysis or sequencing of the hypervariable segments (usually HVS-I) of the control region. A more refined picture about the spread of different mtDNA variants started to emerge during early 1990s with the application of the highresolution restriction fragment analysis to study the mtDNA variation from one continent at a time (Ballinger *et al.* 1992; Torroni *et al.* 1992; 1993*a*; 1993*b*; 1994*a*; 1994*b*; 1994*c*; Chen *et al.* 1995). This series of investigations established "the backbone" of the human mtDNA phylogenetic tree that has largely been in use since then. Meanwhile, an independent classification was proposed, based on the phylogenetic analysis of the variation of HVS-I part of mtDNA (Richards *et al.* 1996). Thereafter, a new series of studies allowed the combination of the control region and coding region (RFLP) data (Torroni *et al.* 1996; Richards *et al.* 1998; Macaulay *et al.* 1999*b*; Schurr *et al.* 1999). As an outcome, it was established that all

mtDNA variants, or *haplotypes*, with characteristic polymorphisms can be divided into a number of monophyletic clades, or *haplogroups* (hgs), defined by single or several mtDNA coding region polymorphisms (often defined by a gain or a loss of a restriction site), associated with polymorphisms (if any) of the mtDNA control region. The main haplogroups are denoted by capital letters and consist of different sub-haplogroups defined by characteristic polymorphisms. Most of the haplogroups exhibit restricted geographical spread, thus distinguishing populations from different continents/geographical regions. The refinement of the mtDNA tree and the nomenclature of mtDNA is an ongoing process. In particular, data emerging from complete mtDNA sequences (Ingman *et al.* 2000; Reidla *et al.* 2003) offer better phylogenetic resolution of different variants of maternal lineages. In the next chapters the global mtDNA tree will be discussed in greater detail.

21. Geographic distribution of mtDNA variants

Owing to the relatively high rate of mutation events within the mtDNA molecule (the molecular clock ticks rapidly) many different variants of mtDNA are found in the human gene pool. They may vary in length, conformation, sequence of the control and coding regions or multiple restriction fragment length polymorphism (RFLP) (Wallace *et al.* 1995). Tracing migrations of *Homo sapiens* with the help of mtDNA is based on the observation that the occurrence of certain haplotypes is often associated with certain world regions and the assumption that this is a result of accumulation of various mutations in different maternal lineages that occurred as people migrated and inhabited new regions. Two main theories exist about the origin and migrations of modern humans. These are expressed as the "Out of Africa model" (Wilson and Cann 1992, Stringer and Mckie 1996) and "Multiregional continuity model" (Thorne and Wolpoff 1992). The first one is more widely accepted due to strong genetic evidence. This theory arose on the basis of the degree of genetic variation patterns in modern human populations. In the late 1980s, Cann, Stoneking and Wilson (1987) demonstrated that human contemporary populations are surprisingly homogeneous, which was later confirmed by Ingman *et al.* (2000). It was in apparent contrast to our closest relatives, the chimpanzees (Cavalli-Sforza 2000). Cann *et al.* (1987) has also showed that the highest variation of mtDNA sequences occurs in African populations.

Africa

According to the African exodus theory, as confirmed by historical mtDNA sequence alterations, our most recent common ancestor (MCRA) arose in central Africa. The time at which anatomically modern *Homo sapiens* diverged from other hominids was most probably between 100-200 ka. The first human lineages that dispersed were described as L1, L2 and L3, with the L1a subcluster being the oldest (Watson *et al.* 1997). The subsequent divergences gave rise to groups L1b, L1c and L2. All of them still dominate in sub-Saharan Africa, which is believed to be the aboriginal place of origin of all mitochondrial sequences, as it has the highest diversity of mtDNA across the world (Jorde *et al.* 1998). The next recorded divergence in the mtDNA (59-69 ka) led to the origin of L3, whose derivatives are now found in nearly all populations in Africa (and nowhere else), and in the clades that first migrated out of the continent.

Asia

Cavalli-Sforza *et al.* (1996) suggested two possible ways of the colonization of Asia. Two major migrations from Africa are also marked in the mtDNA (Maca-Meyer *et al.* 2001). The expansion of haplogroup M (30 – 60 ka), which can be traced from Ethiopia through the Arabian Peninsula to India and Eastern Asia, may represent the southern route, which leads from present- day Ethiopia through the Red Sea to Yemen and then through South Asia. It cannot be excluded, however, that the presence of M haplogroup derivatives along the route is due to back migration from Central Asia, since the M diversity is greater in India (Kivisild *et al.* 1999) than in Ethiopia (Quintarna-Murci *et al.* 1999). The northern route, which led through The Middle East, was taken 40-50 ka by a colonizing group that split into three main clusters. One of them comprises haplogroups W, I and N1b now found in Europe, The Middle East and the Caucasus. I and N1b are also present in Egypt and on The Arabian Peninsula. The next group divided into haplogroup X, found mainly in Europe, and A – common in Asia. The third group subdivided into four lineages. The first one gave rise to haplotype B, which is now found in East Asia, Japan and the southern Pacific Archipelagos. The second formed haplogroups J and T, whereas two others, H and V, belong to the third. Their derivatives are now found in Europe, North Africa and Central Asia. The fourth cluster – U – migrated into different parts of the Old World with the highest frequencies of its sub-haplogroups found in India (U2, U7), North

Africa (U6, U3) and Europe (U5). Lineages associated with sub-haplogroup U6 are the first that migrated back to North Africa, probably during the Upper Paleolithic (Maca-Meyer *et al.*2003). During the same period, lineage U5 was the first that migrated from the Middle East to Europe, where it had its major expansion.

Europe

The present major European lineages are U5, H, I, J, K, T, V, W and X (Richards *et al.* 1998). In contrast to data obtained from analysis of nuclear loci, blood groups and protein markers, the mitochondrial DNA sequence differences suggest only one major colonization and dispersal of human groups in Europe. This conclusion is based on the observation that, with one exception, the territory of Sami (Sajantila *et al.* 1995), the mitochondrial gene pool is very similar throughout the continent (Pult *et al.* 1994) and shows almost no structure, even in regions inhabited by isolated populations such as the Basques (Bertranpetit *et al.* 1995). It has therefore been concluded that the present population structure reflects the first Paleolithic expansion (Comas *et al.* 1997) and subsequent founder effects, with only minor changes brought by the Neolithic spread of agriculture (Richards *et al.* 1996) as evidenced by archeological records. This was further confirmed by estimation of the time of emergence of the European haplogroups as 50 ka for the U5 and 11-14 ka for H, V, I, W, T, K. The latter groups are likely to have diverged from U5 when human populations started to disperse from the glacial refugia after the last Ice Age. Haplogroup J, dating back to 8.5 ka, and which encompasses about 16% of all European mtDNA, is most likely the only one brought to Europe by Neolithic farmers. Its subclades J1b and J1a are clearly distributed along two possible arrival routes: one leading through the valleys of central Europe and the other along the coasts of the Mediterranean Sea and the Atlantic Ocean (Sykes 1999).

The New World

Although some early studies on mitochondrial DNA suggested more than one wave of migration into the Americas (Horai *et al.* 1993), more recent studies indicate only one – an early entry of a single, ancestral lineage of Asian origin (Silva *et al.* 2002). The four most common American haplogroups – A, B, C and D, though relatively old (the estimated time of divergence is 29-42 ka), are very similar in their nucleotide diversity, which suggests their common origin (Silva *et al.* 2002). Mitochondrial DNA data

supports the hypothesis of colonization of Americas through Beringia, where the first inhabitants entered the New World. Some of them migrated into the continent through the ice-free corridor of Alberta, while the others found themselves isolated from the first group after the passage became impassable. Research shows that the diversity of mtDNA among the Amerind, Na-Dene and Eskimo populations is very similar, and that they may be descendents of this population, while Native Americans appear to relate to the Chukchi. Both groups are most likely to have originated from the population that migrated out of Beringia (Bonatta and Salzano 1997).

22. Application of mtDNA study:

22.1. mtDNA and evolutionary study

The reconstruction of evolutionary trees from mitochondrial DNA (mtDNA) data is a common tool with which to infer the relationships of living organisms. The wide use of mtDNA stems from the ease of getting new sequence data for a set of orthologus genes and from the availability of many existing mtDNA sequences for a wide array of species. Though much progress has been made in understanding the parameters that shape the evolution of mitochondria and mtDNA, many questions still remain, and a better understanding of the role this organelle plays in regulating organismal fitness is becoming increasingly critical for accurate phylogeny reconstruction (Ballard, J.W.O. and Rand, D.M. 2005).

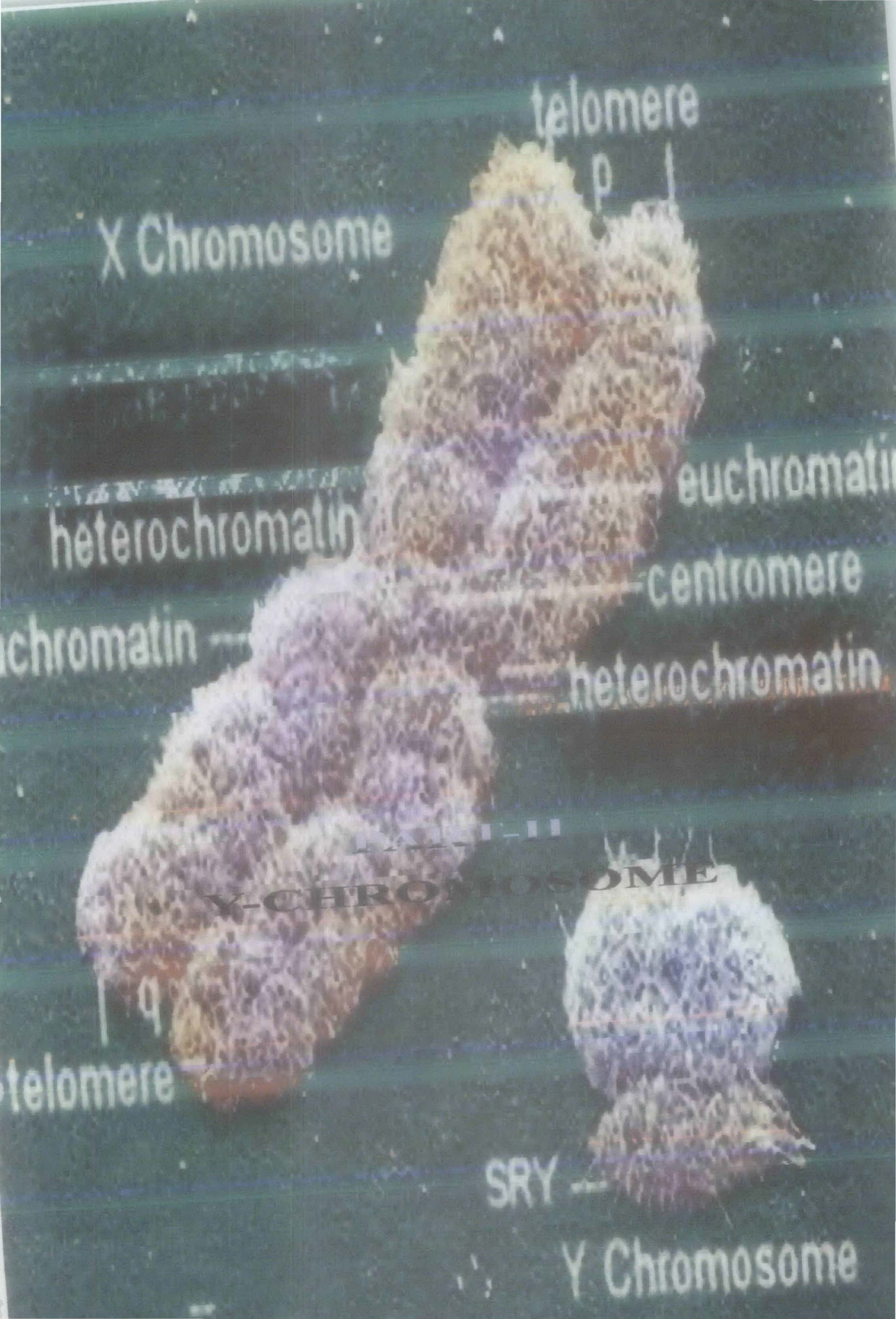
22.2. mtDNA and forensic science

Mitochondrial DNA typing is a method used by forensics scientists to match DNA from an unknown sample to a sample collected at a crime scene. It is ideally used in special cases where the DNA is degraded or the source of the sample doesn't contain enough genomic nuclear DNA for analysis. As it is maternally inherited, the DNA from siblings and all maternal relatives should be identical (in the absence of spontaneous mutations). For this reason, the remains of missing persons can be rapidly identified by using mitochondrial DNA analysis of relatives. Additionally, there is generally a lack of recombination, an event that takes place during nuclear DNA cell division in which two stands of DNA cross over and exchange information, thereby creating greater sequence diversity. Therefore, even matriarchal relatives separated by several generations can serve

as reference samples. Nuclear DNA samples cannot provide this function, due to multiple recombination events that take place throughout the nuclear DNA genome.

22.3. Ethical considerations associated with mtGenome sequencing

There are more than one hundred characterized mtDNA genetic diseases either associated or implicated with mtDNA mutations (Kogelnik *et al.*, 1998). The MitoMap web site contains an expanding list of disease-associated/implicated mtDNA mutations (http://www.mitomap.org/). Given the pivotal role of mtDNA function in basic physiology, the sequencing of genes or gene fragments to increase forensic discrimination cannot confidently avoid encountering a variant that now, or may in the future, be associated with a disease or medically relevant condition. Therefore, unless one focuses on non-phenotypic changes in proteins, ribosomes, or tRNAs of the mtGenome, the forensic scientist could be faced with moral and ethical issues generally associated with clinical genetics and genetic counseling. For example, the group of forensic scientists in Europe focused on sequencing short fragments (50+ nucleotides) of the coding region is using the Pyrosequencing system (Pyrosequencing - Uppsala, Sweden) to increase forensic discrimination (Andreasson *et al.* 2002). One fragment in their assay occurs from the nucleotides 3216-3403. However, within this fragment resides the 3243A-G mutation, a site correlated to three diseases: diabetes mellitus (DM), diabetes mellitus with deafness (DMDF), and mitochondrial encephalomyopathy, lactic acidosis and stoke-like episodes (MELAS). If, having discovered this mutation as part of a family reference sequence to identify a set of skeletal remains, is the forensic scientist now obligated to tell the individual who gave the reference sample and/or other maternal relatives? Conversely, is there a practical means to ensure that the donor or other parties would not discover this information and its medical significance?


telomere
p
X Chromosome
euchromatin
heterochromatin
centromere
chromatin
heterochromatin
Y-CHROMOSOME
q
telomere
SRY
Y Chromosome

Part-II:

1. Why Y?

We inherit many strange things from our parents, biological, behavioural and material, but the males among us inherit from our fathers perhaps the strangest thing of all – a Y chromosome. Ninety-eight per cent of our genome is diploid, with the chromosomes coming in matching pairs, and these pairs engage in recombination, the reshuffling of segments when eggs and sperm are made. The Y chromosome is different: specialised for male sex-determination, it stands aloof and haploid, and, for most of its length, fastidiously avoids the messy business of recombination. This is why the Y chromosome is so widely used in human population studies. As it passes from father to son down paternal lineages, the only changes occurring are due to mutation, and the molecular record of the past is therefore relatively easy to interpret compared to the recombining X chromosome and autosomes. The human Y chromosome determines maleness by causing the development of the testis. It is an unusual segment of the human genome since, apart from two small regions in which pairing and exchange take place with the X chromosome, it is male-specific and haploid, and escapes from recombination. These unique properties of the Y have important consequences for its mutation processes, its genes, and its population genetics: Y chromosomes pass down from father to son largely unchanged, except by the gradual accumulation of mutations. By examining the differences between modern Y chromosomes (as DNA polymorphisms) we can attempt to reconstruct a history of human paternal lineages (Jobling *et al.* 1995). This complements maternal lineage studies using mitochondrial DNA and studies using biparentally inherited markers in the rest of the genome.

1.1. Sex determination

Sex determination mechanisms have evolved many times among eukaryotes and include a variety of environmental and genetic systems. The evolution of sex itself represents one of the major questions in evolutionary biology. Advantages include, for example, an increased opportunity to adapt to environmental changes due to favorable recombinant types. Recombination can also bunch together several deleterious alleles and eliminate them simultaneously (Crow 1994). Sexual populations can have a more rapid rate of

evolution than an equivalent group of asexual organisms (Barton and Charlesworth 1998). Asexual reproduction on the other hand produces twice the number of offspring compared to a sexually reproducing individual. Hermaphrodites (being both male and female at the same time) combine the two systems of asexual and sexual reproduction. Organisms with male heterogamety, like mammals, are characterized by XY sex chromosomes in males and XX sex chromosomes in females. Sex chromosome aneuploidi, in humans and many other mammals, show that the Y chromosome determines sex since XXY karyotypes provide male phenotype and X0 female phenotype (Jacobs and Strongs 1959). The gene that determines sex in most mammals is *SRY* (sex determining region of the Y).

1.2. Y- chromosome evolution

The mammalian X and Y chromosomes are thought to have evolved from an ordinary pair of autosomes (Graves and Schmidt 1992). Support for this hypothesis, and a proposed 300-million-year timeline for sex chromosome evolution, have emerged from studies of modern X–Y gene pairs (Lahn and Page 1999). Lahn and Page (1999) reasoned that X–Y differentiation would have begun only after X–Y crossing over ceased and the first event, which marked the beginnings of X-Y differentiation, occurred about 240 to 320 million years ago, shortly after divergence of the mammalian and avian lineages. Among the 19 X–Y gene pairs studied, age increased in a stepwise fashion along the length of the X chromosome, in four 'evolutionary strata'. This suggested that at least four events had punctuated sex chromosome evolution, with each event suppressing X–Y crossing over in one stratum without grossly disturbing gene order in the X chromosome. Compared to previous estimates the results of Skaletsky *et al.* (2003) and Rozen *et al.* (2003) give the time scale that extends from approximately 4 million years for the X-transposed sequences — the youngest known sequences in the NRY, to approximately 300 million years for SRY — the sex determinant and arguably the oldest gene in the NRY. All NRY X-degenerate genes and pseudogenes seem to be products of a single molecular evolutionary process: the region-by-region suppression of crossing over in ancestral autosomes, with subsequent differentiation of the Y from the X chromosome (Lahn and Page 1999). At least two of the NRY's ampliconic gene families are thought to be also originated in this manner, but subsequently acquired the characteristics of

ampliconic sequences. Inversions in the Y chromosome may have suppressed crossing over with the X chromosome. The findings about the X-degenerate regions support the theory of the genetic benefits of sexual recombination through meiotic crossing over, and the deleterious consequences of its absence (Skaletsky *et al.* 2003). According to this theory, most ancestral genes remained functionally intact in the X chromosome, where the benefits of crossing over (in females) continued. In the Y chromosome, in contrast, the shutting down of X–Y crossing over during evolution triggered a monotonic decline in gene function. In this light, the protein-coding genes in the modern NRY's X-degenerate sequences appear as rare examples of persistence in the absence of sexual recombination. The situation is different for ampliconic genes. Despite the wide variety of genomic sources and mechanisms that gave rise to the ampliconic genes (Ska- letsky *et al.* 2003), they came to exist in the NRY in multiple, nearly identical copies in palindromes. Therefore, the understanding, how the ampliconic genes avoid degradation due to mutation in the absence of crossing-over may be an important outcome from the sequencing of NRY. The possible mechanism that preserves the genes in palindromes is likely to be gene conversion by which the Y chromosome repairs mutations that occur within these genes (Rozen *et al.* 2003). The occurrence of NRY gene pairs that are subject to frequent gene conversion might provide a mechanism for conserving gene functions across evolutionary time in the absence of crossing-over.

2. Structure of Y-Chromosome

Humans, as all other mammals, have an XY chromosomal sex determining system, in which a small Y chromosome triggers male development, and contains genes required for spermatogenesis. Genomic studies have revealed that 95% of Y chromosome's length comprises a region with no X-Y crossing over in male meiosis. This region has come to be named the non-recombining region (NRY) or male-specific region (MSY) (Skaletsky *et al.* 2003). The NRY is flanked on both sides by pseudoautosomal regions, where X-Y crossing over is regular event in male meiosis (Cooke *et al.* 1985; Simmler *et al.* 1985; Skaletsky *et al.* 2003), indicating that however different the X and Y chromosomes are today in size and gene content, they were once homologous (Delbridge and Graves 1999). Human Y-chromosome is about 63 million base pairs long, comprising roughly of 23 Mb of euchromatic DNA, including 8 Mb on the short arm (Yp) and 14.5 Mb on the

long arm (Yq) **(Figure 3).** The heterochromatic region of Y-chromosome consists of three heterochromatic blocks. One heterochromatic sequence is situated in the centromeric region; second, much longer heterochromatic block spans the bulk of the distal long arm. The third block interrupts the euchromatic sequence on Yq **(Figure 3)** (Skaletsky *et al.* 2003). NRY contains 156 known transcription units; all of them are located in the euchromatic part of the Y-chromosome. 78 transcription units encode at least 27 distinct proteins or protein families, the remaining 78 lack strong evidence of protein coding. Of the NRY's 27 distinct protein-coding genes or gene families identified so far, 12 are expressed throughout the body and 11 are expressed exclusively or predominantly in testes (Lahn and Page 1997; Skaletsky *et al.* 2003). Nearly all euchromatic sequences can be divided into three classes: X-transposed, X-degenerate and ampliconic. X-transposed sequences are 99% identical to DNA sequences in Xq21, their presence in NRY is the result of massive X-Y transposition that occurred about 3-4 million years ago (Mumm *et al.* 1997; Schwartz *et al.* 1998). Subsequently, an inversion within the NRY short arm cleaved the X-transposed block into two non-contiguous segments with the combined length of 3.4 Mb. Numerous interspersed repeat elements and only two genes are identified within these segments (Skaletsky *et al.* 2003). X-degenerate sequences contain single-copy gene or pseudogene homologues of 27 different X-linked genes, 16 of which encode NRY's distinct proteins or protein families. These singe-copy genes or pseudogenes display between 60% and 96% nucleotide sequence similarity to their X-linked homologues, and they seem to be surviving relics of ancient autosomes from which the X and Y chromosomes coevolved. 12 of the 16 genes encoded in the X-degenerate sequences are expressed throughout the body (Skaletsky *et al.* 2003). The third class of euchromatic sequences, the ampliconic segments, exhibit as much as 99,9% identity over tens or hundreds of kilobases to other sequences in the NRY. The sequences are located in seven segments that are scattered across Y-chromosome's long arm and short arm, and whose combined length is 10.2 Mb. The most pronounced structural features of ampliconic regions are eight long highly symmetrical palindromes. In addition, the ampliconic regions contain five sets of more widely spaced inverted repeats and a variety of long tandem arrays. Including both coding and non-coding elements, the ampliconic sequences hold 135 of 156 NRY identified transcription units and are

predominantly or exclusively expressed in testes (Skaletsky *et al.* 2003).

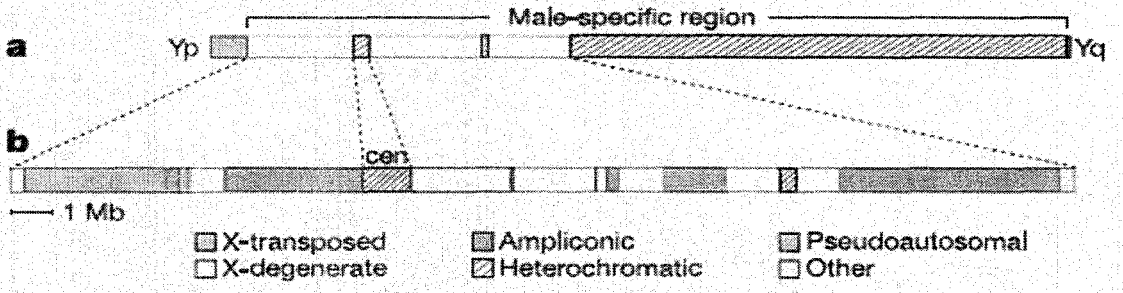


Figure 3. a) A schematic representation of the whole Y-chromosome. **b)** Enlarged view of a 24-Mb portion of the NRY including all three classes of euchromatic sequences. A 1-Mb bar indicates the scale of the diagram (Skaletsky *et al.* 2003).

2.1. Genes of Y-chromosome

Compared to the other human chromosomes, theY chromosome has a limited number of genes. The Y gene poverty may have been the result of the known the tendency of Y chromosome's genes to degenerate during evolution, being nowadays the relic of an ancient common ancestry with the X chromosome (Graves 1995). Both mammalian X and Y chromosomes evolved from ancestral autosomes. The most ancestral gene functions were retained on the nascent X chromosome but deteriorated on NRY portion of the emerging Y (Bull JJ)g iving females with two copies but males with only one copy of many genes.

In Y chromosome not including pseudoautosomal genes, genes include:

- NRY, with corresponding gene on X chromosome
  - AMELY/AMELX (amelogenin)
  - RPS4Y1/RPS4Y2/RPS4X (Ribosomal protein S4)

- NRY, other
  - AZF1 (azoospermia factor 1)
  - BPY2 (basic protein on the Y chromosome)
  - DAZ1 (deleted in azoospermia)
  - DAZ2
  - PRKY (protein kinase, Y-linked)
  - RBMY1A1
  - SRY (sex-determining region)
  - TSPY (testis-specific protein)

- USP9Y
- UTY (ubiquitously transcribed TPR gene on Y chromosome)
- ZFY (zinc finger protein)

(http://en.wikipedia.org)

2.2. Non recombining region of the Y- Chromosome (NRY)

Non recombining region of the Y chromosome splits roughly into two large parts: euchro- matic and heterochromatic portions.

2.2.1. Heterochromatic region of NRY

Efforts to gain sequence-based understanding of human chromosomes have largely by-passed heterochromatic regions (Venter *et al.* 2001). In addition to earlier known centromeric heterochromatin (Tyler-Smith *et al.* 1993) and much longer heterochromatic block (roughly 40 Mb) that comprises the bulk of the distal long arm, a third heterochromatic block — a sharply demarcated island that spans approximately 400 kb, comprises 3,000 tandem repeats of 125 base pairs (bp) and interrupts the euchromatic sequences of proximal long arm of Y chromosome (Yq) — was discovered and characterized by Skaletsky *et al.* (2003), see fig. 1. It was found, that the heterochromatin of NRY encompasses at least six distinct sequence classes, each of which form long, homogeneous tandem arrays (Skaletsky *et al.* 2003).

2.2.2. Euchromatic region of NRY and its characteristic sequence classes

The euchromatic DNA sequences total roughly 24 Mb, including 8Mb on the short arm (Yp) and 14.5Mb on the long arm (Yq), with two minor gaps **(Fig. 3B)**. Nearly all of the euchromatic sequences fall into three distinctive classes — X-transposed, X-degenerated and ampliconic segments. The presence of X-transposed sequences in the human MSY is a result of a massive X-to-Y transposition that occurred about 3–4 million years ago, after the divergence of the human and chimpanzee lineages (Rozen *et al.* 2003) and they are 99% identical to DNA sequences on the long arm of X chromosome (Xq21). Subsequently, an inversion within the NRY short arm cleaved the X- transposed block into two non-contiguous segments, as observed in the modern human NRY (fig. 1B). The X-transposed sequences do not participate in X–Y crossing over during male meiosis, distinguishing them from the pseudoauto- somal sequences. Within the X-transposed segments, which have a combined length of 3.4Mb, only two genes were identified, both

of which have homo- logues on Xq21. Thus, the X-transposed sequences exhibit the lowest density of genes among the three sequence classes in the NRY euchromatin, as well as the highest density of interspersed repeat elements (Skaletsky *et al.* 2003). In contrast to X-transposed regions, the X-degenerate segments of the NRY are dotted with single-copy gene or pseudogene homologues of 27 different X- linked genes. These single-copy NRY genes and pseudogenes display between 60% and 96% nucleotide sequence identity to their X-linked homologues, and they seem to be surviving relics of ancient autosomes from which the X and Y hromosomes co-evolved (Lahn and Page 1999). In 13 cases, the NRY homo- logue is a pseudogene, in the remaining cases, the NRY homologue seems to be a transcribed, functional gene, and the X- and Y-linked genes encode very similar but non-identical protein isoforms. Notably, all 12 ubiquitously ex- pressed NRY genes reside in the X-degenerate regions. Conversely, among the 11 NRY genes expressed predominantly in testes, only one gene, the sex-determining region (SRY), is X-degenerate (Skaletsky *et al.* 2003). The third class, ampliconic segments or segmentally duplicated portion of NRY- name introduced by Hurles and Jobling (2003) — are composed largely of sequences that exhibit marked similarity: as much as 99.9% identity to other sequences in the NRY (Skaletsky *et al.* 2003). The amplicons are located in seven segments that are scattered across the euchromatic long arm and the proximal part of the short arm **(Fig. 3B)** with their combined length of 10.2Mb. The ampliconic sequences exhibit the highest density of genes among the three sequence classes in the NRY euchromatin. Nine distinct NRY-specific protein- coding gene families were identified, with copy numbers ranging from two to approximately 35. All protein-coding families in the ampliconic regions are expressed predominantly or exclusively in testes. The most pronounced structural features of the ampliconic regions are eight massive palindromes, with arm-to-arm nucleotide identities of 99.94–99.997%. The eight palindromes collectively comprise 5.7Mb, or one quarter of the NRY euchromatin. Six of the eight palindromes carry recognized protein-coding genes, all of which seem to be expressed specifically in testes (Skaletsky *et al.* 2003). In all known cases of genes on NRY palindromes, identical or nearly identical gene copies exist on opposite arms of the palindrome. In addition to the palindromes, the ampliconic regions contain five sets of more widely spaced inverted

repeats (IRs). Three of these (IR1, IR2 and IR3) exhibit nucleotide identities of 99.66–99.95%. The ampliconic regions contain also a variety of long tandem arrays.

3. Biological functions of Y-chromosome

Several phenotypes have been associated with the nonrecombining portion of the Y chromosome. For obvious reasons, most of these are male-specific and make the Y a specialised chromosome during human evolution. The most characterising features of this chromosome remain its implication in human sex determination and in male germ cell development and maintenance.

3.1. SRY gene and sex determination

In 1990, the gene responsible for testicular determination, named SRY (Sex-determining Region on theY chromosome), was finally identified (Sinclair *et al.* 1990). *SRY* was cloned by isolation of small fragments of translocated Y on XX sexreversed patients. This gene is located on the short arm of the Y chromosome close to the pseudoautosomal boundary. It comprises a single exon encoding a protein of 204 amino acids which presents conserved DNA-binding domain (the HMG-box: High Mobility Group), suggesting this protein regulates gene expression. This gene has been shown to be essential for initiating testis development and the differentiation of the indifferent, bipotential, gonad into the testicular pathway. Moreover, *SRY* has been proposed to be the master gene regulating the cascade of testis determination. Although many genes and loci have been proposed to interact with *SRY* protein, such as*WT-1* (Wilm's tumour gene), *SF-1* (Steroidogenic Factor 1) and *SOX-9*, the question of how these genes are regulated, if so, by SRY is still unanswered.

3.2. Anti-Turner syndrome effect

Turner syndrome is characterised by a female 45 X karyotype or monosomy X. The principal manifestations of this syndrome are growth failure, infertility, anatomic abnormalities, and selective cognitive deficits. This human genetic disorder is ascribed to haplo-insufficiency of genes of theXchromosome that are common to both X and Y. These genesmust escape X-inactivation because otherwise no difference will be observed between 45, X and 46, XX females. Secondly, in 46, XY these genes must have a male counterpart on the Y responsible to simulate the effects of their X homologues. Although there is no formal identification of genes involved in Turner syndrome, there appear to be

different loci on the X and Y chromosome associated with Turner characteristic features, such as *SHOX/PHOG* (Rao *et al.* 1997, Ellison *et al.* 1996), *ZFX/ZFY* (Page *et al.* 1988), *GCY* and *TCY* (Barbaux *et al.* 1995).

3.3. Oncogenic role of the Y chromosome

The implication of the Y chromosome in cancer remains still speculative.Y chromosome loss and rearrangements have been associated with different types of cancer, such as bladder cancer (Sauter *et al.*1995), male sex cord stroma tumours (de Graaff *et al.* 1999), lung cancer (Center *et al.*1993) and esophageal carcinoma (Hunter *et al.* 1993). Although loss and rearrangements of this chromosome are relatively frequent in different types of cancer, there is no direct evidence for a role of Y in tumour progression since no proto-oncogenes, tumour suppresser genes or mismatch repair genes have been localised to the Y chromosome.

3.4. Genes controlling spermatogenesis

Many of the AZF genes have been proposed as candidate genes involved in human male fertility on the basis of their expression profiles (testis-specific or highly expressed in testis) and sterile phenotypes from targeted disruption of their homologues in mice. However, no direct relation between a Y chromosome gene and male infertility has been demonstrated. In a recent paper, Page and coworkers (Sun *et al.*1999) relate spematogenic failure to a single mutation in a Y-linked gene in AZFa: the USP9Y or, also called, DFFRY. They found a *de novo* 4 bp deletion in a splice-donor site of this gene present in a patient with nonobstructive azoospermia but absent in his fertile brother. This mutation causes protein truncation leading to spermatogenic arrest. Such findings suggest Y-linked USP9Y gene may has a role in human spermatogenesis.

4. Y- Chromosome diversity

Intraspecific diversity as well as interspecific divergence can be estimated by pi (π) or theta (θ). Theta is the proportion of nucleotide sites that are polymorphic in a sample. Pi is the average number of nucleotide differences per site between two randomly chosen sequences in the sample. While θ is a measure of nucleotide polymorphism in a sample and can be corrected for by sequence length and sample size, measures the nucleotide diversity with regard to frequencies of different alleles. Under neutrality, these estimates should be the same but selection and population structure will affect the estimates of π

and θ in different ways (Li 1997). Levels of genetic variation in the Y chromosome may differ from that of the rest of the genome for a number of reasons discussed below.

4.1. Effective population size

In sexual populations, half of the alleles are derived from females and half from males. The number of chromosome variants maintained in the population is, among other things, dependent on the effective population size (Ne) of each chromosome. In an ideal population the relationship between Y, X and autosomes is Ne:3Ne:4Ne suggesting an expected 1:3:4 relationship in diversity between the different chromosomes.

4.1.1. Mating systems

Different mating systems can cause differences in effective population size. Skewed mating systems, for example in polygynous species, where one male mates with many females, will affect the relative difference in effective population size between chromosomes. For example: the Y:X:autosome relationship when one male mates with two females will be 1:5:6. If the ratio of females to males is increased to ten (which is common in some species (Roed *et al.* 2002) the relationship would be 1:21:22.

4.2. Selection

New mutations can be neutral, advantageous or disadvantageous. The probability of fixation or elimination of the mutation in the population depends on the relative fitness of the new phenotype. Exceptions occur for balancing selection, over dominance (where heterozygotes are favored) and in limited populations. Negative selection will tend to eliminate disadvantageous mutants or genotypes from the population and is the prevailing type of selection since the majority of non-neutral mutations are deleterious or slightly deleterious. Positive selection increases the probability for an advantageous mutation to become fixed in the population (Li 1997). However, the chance of losing a new advantageous mutation from the population by random genetic drift (change in allele frequency due to chance) can still be high (Hartl and Clark 1997).

Selection at a locus will also affect linked sites. In the absence of recombination, selection will tend to reduce genetic variability at linked sites to the same extent as at the locus under selection. With recombination, the effect becomes gradually smaller as the rate of recombination between the selected locus and linked sites increases. In line with this thinking, levels of neutral variability have been shown to correlate with

recombination rate in humans (Nachman 2001), mice (Nachman 1997), plants (Stephan and Langley 1998) and fruit flies (Begun and Aquadro 1992). The Y chromosome, which lacks recombination (except in the PAR), should be expected to have reduced variation as compared to recombining chromosomes. Selective sweeps and background selection may have severe affects on the MSY where all sites are linked compared to other genomic regions.

4.2.1 Selection sweeps

Selective sweep or the hitchhiking effect is an effect of positive selection where a favorable allele drives through the population to fixation together with its' linked loci. This will reduce the variation linked to the selected site and decrease the diversity in the population (Rice 1987). The impact of a selective sweep depends on the recombination rate and selection coefficient, the lower the recombination rate and/or the higher the selection coefficient, the larger is the genomic region affected by the sweep. In the Y chromosome where 95% of the sites are linked, an advantageous gene regulating a male specific trait, like one involved in spermatogenesis, may sweep through the population and eliminate all variation in the MSY (Roldan and Gomendio 1999; Wyckoff *et al.* 2000). Selective sweeps can bring about fixed Y chromosomes within a species and different between species, while mutations will only slowly produce new variants in a population. In contrast to the neutralists prediction of a positive correlation of intraspecific variation and interspecific divergence, positive selection can lead to uncoupling of levels of polymorphism and divergence (Li 1997).

4.2.2. Background selection

Background selection is an effect of negative selection where deleterious mutations will be eliminated from the population together with their linked loci. This process, as with selective sweeps, will reduce variation in the region around the selected site (Figure 10). Similar to sweeps the impact of background selection depends on the recombination rate and the selection coefficient. Background selection is not thought to alter allele frequencies to the same extent as selective sweeps; indicating that the two types of selection can be distinguished from each other (Charlesworth *et al.* 1995). In a non-recombining region, mildly deleterious as well as weakly advantageous alleles will survive linked to each other. In the absence of a strongly advantageous mutation, a

neutral or weakly selected mutation can only survive on a non-recombining chromosome (like Y) if there is no strongly deleterious mutation, otherwise it will be eliminated (Charlesworth, 1994).

4.3. Sex specific mutation rates

Mutations are generated during DNA replication. The number of germ cell divisions differs between spermatogenesis and oogenesis. In oogenesis every mature oocyte has gone through a total of 24 cell divisions irrespective of the age of the female. In spermatogenesis however cell division is a continuous process, so the older the male the more cell divisions his sperms have gone through. For example in a 20 year-old man every sperm has gone through about 150 cell divisions and at the age of 40, 610 cell divisions (Hurst and Ellegren1998). The male to female mutation rate ratio, α_m, is mostly dependent on the skewed number of cell division in the germ lines of males and females. As α_m is generally larger than one, meaning that male germ cells mutate more frequently than female germ cells, more mutations in the Y chromosome than in other chromosomes can be predicted (Miyata *et al.* 1987).

4.4. Other factors affecting Y- diversity

Differences in migration between males and females can produce variation in the patterns of genetic differentiation detected in maternally and paternally inherited systems. In a patrilocal species, where female migrate more than males, this will imply less variation in the Y chromosome locally. In a global perspective, this will lead to higher differentiation in Y chromosome than in the maternally inherited mitochondrial DNA (Seielstad *et al.* 1998). Spermatogenesis and sperm mobility are energy demanding processes;therefore, the function of the mitochondria is vital for reproductive success. Deleterious mutations in the mtDNA that affect the energy production negatively would be expected to lead to impaired reproduction and, consequently, reduced effective population size among males. This will lower the effective population size of Y chromosomes and reduce its diversity (Gemmell and Sin 2002).

4.5. Mutation and Y-chromosome diversity

As mutation is the only force that acts to diversify Y haplotypes, understanding mutational dynamics is important to understanding the origins of haplotype diversity.More generally, knowledge of the rates and processes of mutation at different classes of marker is

fundamental to evolutionary interpretations of diversity data and to understanding genetic disease. In this broader context, studies on the Y chromosome are of particular interest because the mutations observed here are the results of exclusively intra-allelic processes. Although these processes occur on all chromosomes, the haploid Y provides a model for studying them without the complicating factors of inter-allelic events and allelic diversity. (Mark A. Jobling and Chris Tyler-Smith 2003)

5. The development of Y- markers

Evolutionary or population genetic studies focused at male-specific patterns obviously require genetic markers from the Y chromosome. This may be in the form of polymorphic sequences for intraspecific studies, but may in principle concern any Y chromosome specific sequence. In human and mouse the genome projects have been advanced producing a lot of Y chromosome sequence for these species (International Human Genome Sequencing Consortium 2002; Skaletsky *et al.* 2003; Waterston *et al.* 2002). In the absence of large-scale genome sequence information, alternative approaches are needed. The two most popular methods are random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP). RAPD is a very fast method with low cost (Hadrys *et al.* 1992) but hard to reproduce since the conditions must be exactly the same from time to time. AFLP on the other hand is very consistent but time consuming and more expensive than RAPD (Mueller and Wolfenbarger 1999). Both methods can be used with an almost unlimited set of primers that will render an inexhaustible source of possibilities to develop Y markers.

5.1. Random amplified polymorphic DNA (RAPD)

Random amplified polymorphic DNA-polymerase chain reaction (RAPD-PCR) uses a single short primer, which will anneal to various places in the genome and amplify fragments of different length in the PCR (Williams *et al.* 1990). These fragments can be separated on an agarose gel and male- specific bands can be isolated and sequenced to develop locus-specific Y-linked markers (Gutierrez-Adan *et al.* 1997; Olivier and Lust 1998; Wardell *et al.* 1993).

5.2. Amplified fragment length polymorphism (AFLP)

Double stranded adapters are ligated onto digested DNA. These adapters are complementary to the primers used in the next step, which is a preselective amplification

step. The complementary primers have an extra base at the 3′end (P-n), which extends into the genomic DNA, amplifying only one fourth of the fragments. After this, a second amplification step is performed with primers similar to the preselective one but with two additional bases at the 3′ end (P-nnn) amplifying only 1/16 of the fragments from the preselective PCR. The results can then be displayed on a denaturing acrylamide gel either by silver staining or by radioactive labeled primers (Vos *et al.* 1995).

5.3. Other techniques

There are other methods that can be used to get Y specific sequence, for example reduced representation shotgun (RRS) (Altshuler *et al.* 2000) together with flow sorting (Mullikin et al. 2000). The most widely used methods for targeted SNP discovery are those exploiting the difference between the mismatched heteroduplex DNA from the perfectly matched homoduplex DNA (Kwok and Chen 2003). Some methods like Single Strand Conformation Polymorphism (SSCP) distinguish between the three dimensional structure of single-stranded DNA held in constant environment. Regardless of the mechanism of variation scanning, a separation step by gel or capillary electrophoresis or high-performance liquid chromatography is required (Kwok 2001). The most extensively employed application for local SNP discovery in case of Y chromosome is denaturing high performance liquid chromatography (DHPLC) (Underhill *et al.* 1997; Underhill *et al.* 2000; Thomson *et al.* 2000; Shen *et al.* 2000). Recent choice of Y-chromosome analysis is by multiplex SNaPshot approach or SNP stream Genotyping System. This provides an automated scalable system capable of performing from 4,600 to over 3,000,000 SNP genotypes per day.

6. Y- Chromosome as a tool to study human genetic variation and demographic history

Until recently, the Y chromosome seemed to fulfil the role of juvenile delinquent among human chromosomes — rich in junk, poor in useful attributes, reluctant to socialize with its neighbours and with an inescapable tendency to degenerate. The availability of the near-complete chromosome sequence, plus many new polymorphisms, a highly resolved phylogeny and insights into its mutation processes, now provide new avenues for investigating human evolution (Mark A. Jobling and Chris Tyler-Smith 2003). The Y chromosome contains the largest non-recombining block in the human genome. By virtue

of its many polymorphisms it is now the most informative haplotyping system, with applications in evolutionary studies, forensics, medical genetics, and genealogical reconstruction. However, the emergence of several unrelated and non-systematic nomenclatures for Y-chromosomal binary haplogroups is an increasing source of confusion.

6.1. Special features of Y-Chromosome

The unique properties of Y-chromosome make it a suitable tool for investigating phylogenetics and human evolution (Jobling and Tyler-Smith 2003), DNA forensics (Jobling *et al.* 1997), medical genetics (Jobling and Tyler- Smith 2000) and for genealogical reconstruction (Jobling 2001). Y-chromosome is inherited from father to son and as 95% of it does not recombine in male meiosis (Skaletsky *et al.* 2003), the combination of alleles passes intact from one generation to the next, changing only by mutation. Because of its haploidity and male specificity, effective population size of the Y-chromosome is one-quarter of any autosome and onethird of the X-chromosome, assuming equal numbers of males and females in the population. As the result, Y-chromosome is more susceptible to genetic drift (Underhill *et al.* 2001). In addition, social behavior of men, who are the carriers of Y-chromosome, affects the geographical clustering of Y-chromosome lineages. Since in most societies men tend to live closer to their parents than women, the local clustering and differentiation of Y variants is enhanced (Seielstad *et al.* 1998). These distinctive features make Y-chromosome a useful instrument for studying past migrations and population events.

6.2. Types of Y-Chromosome markers

Y chromosome variation consists of large amount of different types of poly- morphisms, which are widely used in evolutionary studies. They may roughly be divided into two large groups: bi-allelic markers and polymorphisms of tandem repeats or multi-allelic markers (Jobling and Tyler-Smith 2000). Biallelic markers include SNPs (Single Nucleotide Polymorphisms) and insertions and deletions (indels). SNPs are the most common type of poly- morphisms, constituting more than 90% of total polymorhpisms of DNA (Collins *et al.* 1997). Only these bi-allelic mutations that have occurred, highly likely, only once in history of humans and have a detectable frequency in human populations are used in phylogenetic studies. Sometimes recurrent mutations are also

used (YCC 2002), but then the recurrent mutation has to have occurred in different, unambiguously independently distinguishable branches in phylogenetic tree of Y chromosome, like SRY 1532 in background of haplogroups A and R1a. SNPs are characterized by low mutation rate and are therefore suitable for studying early demographic events in human history. The mutation rate for SNP markers is considered to be an average on the order of 2×10^{-8} per base per generation (Nachman and Crowell 2000). In the case of indels (YAP insertion, 12f2 deletion), their presence or absence compared to ancestral state is detected. Large rearrangements, mostly deletions in regions of Y-specific genes (AZFa, AZFb, AZFc), have been known as causes for many diseases leading to male infertility, causing spermatogenic failure, azoospermia, severe oligo-spermia or otherwise severely impair male reproductive fitness (Fernandes *et al.* 2002; 2004; Repping *et al.* 2003; 2004). Not all indels affect male fertility: they persist over generations and are sufficiently common to be considered as polymorphisms. One such example is a 2kb deletion in 12f2 marker (Casanova *et al.* 1985), used for defining haplo-group J according to the present nomenclature (YCC 2002). Some indels have arisen independently more than once in human history. For example, the deletion or duplication of the 50f2/C (DYS7C) region in background of different haplogroups is thought to be arisen at least 7–8 times (Jobling *et al.* 1996). Another example is the deletion of DAZ3/DAZ4 region that has been indicated to occur in haplogroup N individuals (Fernandes *et al.* 2004), widely spread in northern Eurasia. These findings show that new informative variations in different Y-chromosomal haplogroup backgrounds in phylo-genetic tree may be detected also in studies, otherwise focused on medical aspects. Another frequent type of polymorphisms, present also in Y chromosome, is tandem repeats, mostly in non-coding DNA regions (Chambers and MacAvoy 2000). According to their length, these repeats are classified as satellite-DNAs (repeat lengths of one to several thousand base pairs), minisatellites or variable number of tandem repeats, (VNTRs) ranging from 10 to 100 bp, and microsatellites or short tandem repeats (STRs), with motifs less than 10 bp, mostly 2 to 6 bp long (Chambers and MacAvoy 2000). In Y-chromosomal studies microsatellites are widely used, while minisatel- lites have been used only in some investigations (Jin *et al.* 2003). Microsatellites are multi-allelic markers with different allele numbers ranging from 3 to 49 in locus (de Knijff *et al.* 1997)

and were first taken in use by Litt and Luty (1989). Their mutation rate is much higher than that for bi- allelic markers and, therefore, they are widely used in phylogenetic studies to investigate details of demographic events that have occurred in a more recent time-scale. In evolutionary studies STRs are valuable in combination with binary haplogroup data (de Knijff 2000), as they enable to study diversity within a haplogroup. STRs are particularly widely explored in forensic work (Jobling *et al.* 1997). So far the number of widely used Y-chromosomal STRs has been quite low (about 30) but in a recent study by Kayser *et al.* (2004), 166 new and potentially useful STRs were described. Different analyses have shown that the average mutation rate for autosomal tetranucleotide repeats is about 2.0×10^{-3} per generation (Weber and Wong 1993). Similar results (2.0×10^{-3} per generation) were obtained also for Y- chromosomal tetranucleotide repeats in deep pedigree studies (Heyer *et al.* 1997) that were in concordance with results of Weber and Wong (1993) for autosomal microsatellites. Mutation rate was studied in father-son pairs, getting average mutation rate 2.8×10^{-3} by Kayser *et al.* (2000b), while in Forster *et al.* (2000), by using the haplotype network of Native Americans the average rate was found to be 2.6×10^{-4} per generation (here: 20 years) per locus, differing thus about an order of magnitude from results by Heyer *et al.* (1997). In Zhivotovky *et al.* (2004), mutation rate was estimated, using data on microsatellite variation within Y chromosome haplogroups defined by SNPs in populations with documented short-term histories (the African Bantu expansion, the divergence of Polynesian populations and the origin of Gypsy populations from Bulgaria), as well as making use of comparative data on worldwide SNP variation, both at autosomal and Y chromosome loci. The estimated mutation rate for an average Y chromosome short-tandem repeat locus was found to be 6.9×10^{-4} per generation (here: 25 years).

6.3. Could the Y-chromosome recombine?

There have been controversies over whether mitochondrial DNA (mtDNA) recombines93; although the Y chromosome (excluding the pseudoautosomal regions) is assumed to be exempt from recombination, are there any circumstances under which it could occur? Recent work (Rozen, S. *et al.* (2003) has shown extensive gene conversion between paralogues on the Y chromosome and as gene conversion represents a form of recombination,

the argument has been made that the Y chromosome is therefore a recombining chromosome.There are rare instances in which segments of the Y chromosome are carried on other chromosomes as asymptomatic translocations(Cooke, H. J. & Noel, B. 1979; Andersson, M. *et al.* 1988).Recombination (or gene conversion) could occur between a segment from one haplogroup and a normal Y chromosome from a different haplogroup.Typing polymorphisms on a resulting recombinant chromosome would show a burst of HOMOPLASY in the phylogeny, corresponding to a set of markers located together in the recombinant segment.As population sample sizes and numbers of markers increase, it seems probable that such rare recombinant chromosomes will eventually be discovered.

6.4. Selection and Y-chromosome diversity

Aside from mutation, selection is a potentially important force in patterning Y-haplotype diversity in populations. The Y chromosome is subject to purifying selection. For example, absence of the Y chromosome (the 45, X karyotype) leads to Turner syndrome, and the loss or inactivation of Y genes can produce an XY female or hermaphrodite phenotype (Berta, P. *et al.* 1990), or male infertility (Sun, C. *et al.* 1999) (for xample, *USP9Y/DFFRY*).

7. Phylogeny of Y-Chromosome haplogroups

Constant discovery of new markers in the human Y-chromosome led to the emergence of several unrelated and nonsystematic nomenclatures for Y-chromosomal haplogroups (Jobling and Tyler-Smith 2000; Semino *et al.* 2000; Underhill *et al.* 2000; Hammer *et al.* 2001). To resolve the issue, the Y Chromosome Consortium (YCC) created a single most parsimonious Y-chromosomal phylogenetic tree of binary haplogroups (YCC 2002), incorporating virtually all previous nomenclatures. The ancestral states of Y-chromosomal binary markers were determined by comparison of homologous regions of the NRY in closely related species like chimpanzees, gorillas and orangutans (Underhill *et al.* 2000; Hammer *et al.* 2001). A hierarchical nomenclature system and a simple set of rules was constructed to unambiguously label the different clades nested within this tree (YCC 2002). Recently, a revised version of Y-chromosome haplogroup tree was published (Karafet *et al.* 2008). New markers and haplogroups were introduced and previous haplogroups significantly rearranged following the rules presented in YCC 2002.

This most up to date Y-chromosome tree contains 311 distinct haplogroups defined by 599 mutations **(Figure 4)** (Karafet *et al.* 2008).

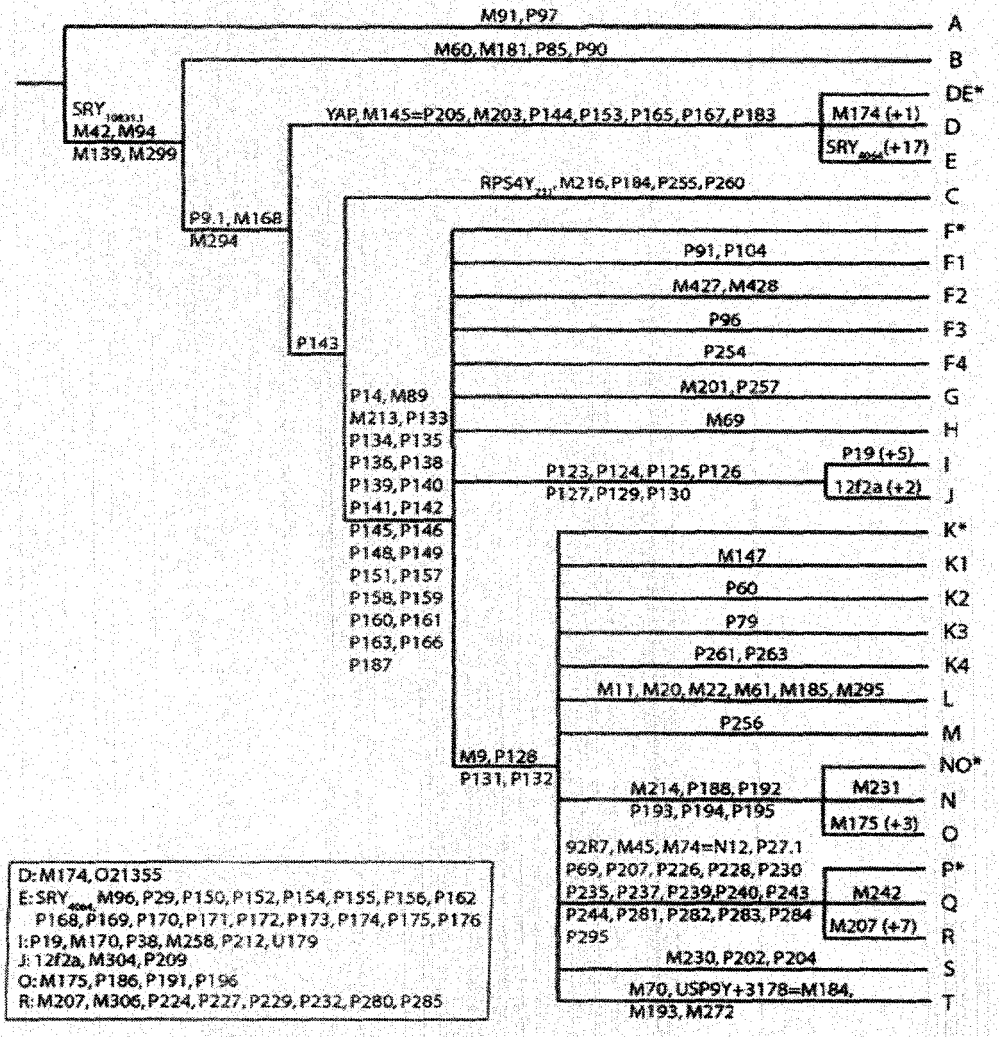


Figure 4. An abbreviated form of the Y-chromosome parsimony tree. Mutation names are indicated on the branches, haplogroup names at the tips of the tree. The symbol * represents potentially paraphyletic lineages called "paragroups" (Karafet *et al.* 2008).

8. Major branches of Y- Chromosomal phylogenetic tree

Groups A and B

Haplogroups A and B are only found in sub-Saharan Africa (and in populations extracted from there in modern times, primarily via Atlantic slave trade and Arabian slave trade).

The first to branch off was A, with defining mutation M91. All other haplogroups are summarized as BT (also referred to as YxA).

- Haplogroup A (M91) Found in Africa, especially the Khoisan, Ethiopians (especially Beta Israel) and Nilotes
- BT (M42, M94, M139, M299) ca. 55 ka BP
    - Haplogroup B (M60) Found in Africa, especially the Pygmies and Hadzabe
    - CT (see below)

Groups with mutation M168 (CT)

The defining mutations separating CT (all haplogroups excepting A and B) are M168 and M294. These mutations predate the "Out of Africa" migration. The defining mutations of DE probably occurred in Northeastern Africa some 50,000 years ago.The P143 mutation that defines Haplogroup CF may have occurred somewhat earlier, perhaps even as early as 55,000 years ago, after the first Out of Africa migration brought Homo sapiens to the southern coast of Southwest Asia.

- Haplogroup CF (P143) Found outside of Africa, throughout Eurasia, Oceania, and the Americas
    - Haplogroup C (M130, M216) Found in Asia, Oceania, and North America
        - Haplogroup C1 (M8, M105, M131) Found in Japan
        - Haplogroup C2 (M38) Found in Indonesia, New Guinea, Melanesia, Micronesia, and Polynesia
        - Haplogroup C3 (M217, P44) Found throughout Eurasia and North America, but especially among Mongols, Kazakhs, Tungusic peoples, Paleosiberians, and Na-Dené-speaking peoples
        - Haplogroup C4 (M347) Found among the indigenous peoples in Australia
        - Haplogroup C5 (M356) Found in the Indian subcontinent and the Arabian Peninsula
    - Haplogroup F (M89, M213) Found in Southern India, Sri Lanka, Yunnan, Korea
        - GT (see below)
- Haplogroup DE (M1, M145, M203) ca. 65 ka
    - Haplogroup D (M174) Found in Tibet, Japan, the Andaman Islands
        - Haplogroup D1 (M15)
        - Haplogroup D2 (M55, M57, M64.1, M179, P12, P37.1, P41.1 (M359.1), 12f2.2)
        - Haplogroup D3 (P47)
    - Haplogroup E (M40, M96)

        - Haplogroup E1 (P147)
            - Haplogroup E1a (M33, M132) Found in Africa; **formerly E1**
            - Haplogroup E1b (P177)
                - Haplogroup E1b1 (P2, DYS391p); **formerly E3**
                    - Haplogroup E1b1a (M2) Found in Africa; **formerly E3a**
                    - Haplogroup E1b1b (M215) Found in East Africa, North Africa, the Middle East, and Europe (especially in areas near the Mediterranean); **formerly E3b**
        - Haplogroup E2 (M75) Found in Africa

Groups descended from Haplogroup F (GT)

The groups descending from haplogroup F are found in some 90% of the world's population, but almost exclusively outside of sub-Saharan Africa. The mutation of IJ corresponds to a wave of migration out of the Middle East or Western Asia some 45 ka that subsequently spread into Europe (Cro-Magnon). Haplogroup G originated in the Middle East or Caucasus, or perhaps further east as far as Pakistan some 30 ka, and spread to Europe with the Neolithic Revolution. Haplogroup H probably occurred in India some 30-40 ka, and remains prevalent there, spreading westwards in historical times with the Romani migration. Haplogroup K probably originated in southwestern Asia and spread widely to Africa, Eurasia, Australia and the South Pacific.

- Haplogroup G (M201) ca. 30 ka Found in many ethnic groups in Eurasia; most common in the Caucasus, the Iranian plateau, and Anatolia; in Europe mainly in Greece, Italy, Iberia, the Tyrol, as well as Bohemia, Moravia; less than 2% in Britain[citation needed], Norway,[1] and Sweden[2]
    - Haplogroup G1
    - Haplogroup G2
        - Haplogroup G2a
            - Haplogroup G2a1
        - Haplogroup G2b
        - Haplogroup G2c (formerly Haplogroup G5)
            - Haplogroup G2c1
- Haplogroup H (M69) Found mainly in South Asia
    - Haplogroup H1
    - Haplogroup H2
- Haplogroup IJ (P123, P124, P125, P126, P127, P129, P130) ca. 45 ka
    - Haplogroup I (M170, M258) Found in Europe and parts of the Near East

        - Haplogroup I1 (M253) Found mainly in northern Europe
        - Haplogroup I2 (P215) Found mainly in southeast Europe and Sardinia
    - Haplogroup J (M304, S6, S34, S35)
        - Haplogroup J* (minimal distribution)
        - Haplogroup J1 Associated with Northeast Caucasian peoples in Dagestan and Semitic peoples in Mesopotamia, the Levant, the Arabian Peninsula, Ethiopia, and North Africa
        - Haplogroup J2 (M172) Found mainly in the Mediterranean basin (esp. Turkey, Greece, the Balkans, and Italy), the Caucasus, Jews, and the Middle East
- Haplogroup K (M9) Found in New Guinea, Australia
    - LT (see below)
    - Haplogroup K1 Found in the Solomon Islands and Fiji
    - Haplogroup T Found in Africa (mainly Afro-Asiatic-speaking peoples), the Middle East, the Mediterranean, South Asia; formerly known as **Haplogroup K2**
    - Haplogroup K3 Found in two individuals in South Asia
    - Haplogroup K4
    - Haplogroup S Found in New Guinea; formerly known as **Haplogroup K5**
    - Haplogroup K6 Found in Melanesia
    - Haplogroup K7 Found in Melanesia

Groups descended from Haplogroup K (LT)

Haplogroup L is mainly found in South Asia. Haplogroup M is most prevalent in New Guinea and Melanesia. The NO haplogroup appeared ca. 35-40 ka in Central Asia. Haplogroup N probably originated in Mongolia and spread both east into Siberia and west, being the most common group found in Uralic peoples. Haplogroup O is found at its highest frequency in East Asia and Southeast Asia, with lower frequencies in the South Pacific, Central Asia, and South Asia. Haplogroup P gave rise to groups Q and R, and is rarely found in its undifferentiated stage. It probably originated in Central Asia or the Altai region. Haplogroup Q also originated in Central Asia, migrating east to North America.

- Haplogroup L (M20) Found in South Asia, Central Asia, Southwest Asia, the Mediterranean
    - Haplogroup L1
    - Haplogroup L2
    - Haplogroup L3

- Haplogroup MNOPS (rs2033003/M526)
  - Haplogroup M (P256) Found in New Guinea and Melanesia
    - Haplogroup M1
    - Haplogroup M2
    - Haplogroup M3
  - Haplogroup NO (M214) 35-40 kya
    - Haplogroup NO* (minimal distribution)
    - Haplogroup N (M231) Found in northernmost Eurasia, especially among the Uralic peoples
    - Haplogroup O (M175) Found in East Asia, Southeast Asia, the South Pacific
  - Haplogroup P (M45)
    - Haplogroup P* (minimal distribution)
    - Haplogroup Q (MEH2, M242, P36) Found in Siberia and the Americas
    - Haplogroup R (M207, M306) Found in Europe, West Asia, Central Asia, and South Asia
  - Haplogroup S (formerly Haplogroup K5) Found in the highlands of New Guinea

- Haplogroup T (formerly Haplogroup K2) (M184, M70, M193, M272) Found in a significant minority of Sciaccensi, Somalis, Eivissencs, Stilfser, Ethiopians, Fulbe, Egyptians, and Omanis; also found at low frequency throughout the Mediterranean and parts of India
  - Haplogroup T1 (M320)
  - Haplogroup T2 (P77)
  - Haplogroup T3 (L131)

Groups descended from Haplogroup NO (M214)

The NO haplogroup appeared ca. 35-40 ka in Central Asia. Its predecessor, haplogroup MNOPS, is ancestral to a range of haplogroups distributed widely across mainly Eurasia, Oceania, and the Americas, namely the M, N, O, Q, R, and S haplogroups. Haplogroup N possibly originated in eastern Asia and spread both west into Siberia and north, being the most common group found in some Uralic speaking peoples. Haplogroup O is found at its highest frequency in East Asia and Southeast Asia, with lower frequencies in the South Pacific, Central Asia, and South Asia.

- Haplogroup NO (M214) 35-40 ka (minimal distribution)
  - Haplogroup N (M231) Found in northernmost Eurasia, especially among the Uralic peoples
    - Haplogroup N1 (LLY22g)

        - Haplogroup N1a
        - Haplogroup N1b
        - Haplogroup N1c
    - Haplogroup O (M175) Found in East Asia, Southeast Asia, the South Pacific
        - Haplogroup O1 (MSY2.2) Found in eastern and southern China, Taiwan, and Southeast Asia, especially among Austronesian and Kradai peoples
        - Haplogroup O2 (P31, M268)
            - Haplogroup O2a (M95) Found in Japan, southern China, Southeast Asia, and the Indian subcontinent, especially among Austro-Asiatic peoples, Kradai peoples, Malays, and Indonesians
            - Haplogroup O2b (SRY465, M176) Found in Japan, Korea, Manchuria, and Southeast Asia
        - Haplogroup O3 (M122) Found throughout East Asia, Southeast Asia, and Austronesia including Polynesia

Groups descended from Haplogroup P (M45)

Haplogroup P (M45) has two branches. They are Q-M242 and R-M207, which share the common marker M45 in addition to at least 18 other SNPs.

Haplogroup Q

Q is defined by the SNP M242. It is believed to have arisen in Central Asia approximately 35-40 000 years ago. The subclades of Haplogroup Q with their defining mutation(s), according to the 2008 ISOGG tree are provided below. ss4 bp, rs41352448, is not represented in the ISOGG 2008 tree because it is a value for an STR. This low frequency value been found as a novel Q lineage (Q5) in Indian populations

The 2008 ISOGG tree

- Q (M242)
    - Q*
    - Q1 (P36.2)
        - Q1*
        - Q1a (MEH2)
            - Q1a*
            - Q1a1 (M120, M265/N14) Found with low frequency among Dungans, Han Chinese, Hazaras, Japanese, Koreans, and Tibetans[3][4]

        - Q1a2 (M25, M143) Found at low to moderate frequency among some populations of Southwest Asia, Central Asia, and Siberia
        - Q1a3 (M346)
            - Q1a3* Found at low frequency in Pakistan, India, and Tibet
            - Q1a3a (M3) Typical of indigenous peoples of the Americas
                - Q1a3a*
                - Q1a3a1 (M19) Found among some indigenous peoples of South America, such as the Ticuna and the Wayuu[5]
                - Q1a3a2 (M194)
                - Q1a3a3 (M199, P106, P292)
        - Q1a4 (P48)
        - Q1a5 (P89)
        - Q1a6 (M323) Found in a significant minority of Yemeni Jews
    - Q1b (M378) Found at low frequency among samples of Hazara and Sindhis

Haplogroup R

Haplogroup R is defined by the SNP M207. The bulk of Haplogroup R is represented in lineages R1a and R1b. R1a likely originated in the Eurasian Steppes, and is associated with the Kurgan culture and Proto-Indo-European expansion. It is primarily found in Central Asia, South Asia, and the Slavic peoples of Eastern Europe. R1b probably originated in Central Asia. It is the dominant haplogroup of Western Europe and also found sparsely distributed among various peoples of Asia and Africa. Its subclade R1b1b2 (M269) is the haplogroup that is most commonly found among modern European populations, especially those of Western Europe.

- Haplogroup R1 (M173) Found throughout western Eurasia
    - Haplogroup R1a (M17) Found in Central Asia, South Asia, and Central, Northern and Eastern Europe
    - Haplogroup R1b (M343) Found in Western Europe, West Asia, Central Asia, North Africa, and northern Cameroon
- Haplogroup R2 (M124) Found in South Asia, Caucasus, Central Asia, and Eastern Europe (http://en.wikipedia.org/)

9. Distribution of the Y-Chromosomal variation in Asia

Asia is a vast continent where genetic drift has played an important role in shaping the

Y-chromosomal variation, affecting more seriously smaller popu- lations. Therefore, the mean values of haplogroup distribution over the whole Asia, taken alone, are not informative for making conclusions about their phylo-geographic spread: 15 of 18 major haplogroups are present in Asia, but their distribution is highly region-specific (Hammer and Zegura 2002). Here the attention is focused mainly on Indian Sub-continent regions and Y-chromosomal haplogroups that have played a crucial role in modern human migration.

9.1. Colonization of South and East Asia

There are two major models of migration routes of the initial peopling of East Asia. The first one argues that an early southeastern Asian spread via the southern route is a more likely scenario, followed by a northward migration. Some genetic surveys, such as the variation of autosomal microsatellite markers (Chu *et al.* 1998) and Y-chromosomal binary markers (Su *et al.* 1999; Jin and Su 2000; Su and Jin 2000) support this model. This result is concordant mainly with the distribution of major Asian hg O with larger diversity in southern *versus* northern regions and its northern sub-clades being a subset of the southern variation (Su *et al.* 1999). In contrast, the second model suggests a bi- and/or multidirectional route: one migration possibly through Central Asia and one through Southeast Asia (Ding *et al.* 2000; Karafet *et al.* 2001; Wells *et al.* 2001; Tajima *et al.* 2002; Jin *et al.* 2003). In previously named studies the main haplogroups for which the possible northern route is discussed are hg C with its very wide overall Asian distribution, together with hg D, the latter displaying a very specific and restricted distribution, absent or present in marginal frequencies in most populations of Asia, but frequent in Tibetan, Andamanese and Japanese populations.

9.2. Y-Chromosomal haplogroup variation in India

Several studies have argued that, in contrast to the relative uniformity of mtDNA, the Y chromosomes of Indian populations display relatively small genetic distances to those of West Eurasians (Bamsad et al. 2001), linking this finding to hypothetical migrations by Indo-Aryan speakers. Wells *et al.* (Wells et al. 2001) highlighted M17 (R1a) as a potential marker for one such event, as it demonstrates decreasing frequencies from Central Asia toward South India. Departing from the "one haplogroup equals one migration" scenario,

Cordaux *et al.* (2004) defined, heuristically, a package of haplogroups (J2, R1a, R2, and L) to be associated with the migration of IE people and the introduction of the caste system to India, again from Central Asia, because they had been observed at significantly lower proportions in South Indian tribal groups, with the high frequency of R1a among Chenchus of Andhra Pradesh (Kivisild *et al.* 2003) considered as an aberrant phenomenon (Cordex *et al.* 2004). Conversely, haplogroups H, F*, and O2a, which were observed at significantly higher proportions among tribal groups of South India, led the same authors to single them out as having an indigenous Indian origin. Only O3e was envisaged as originating (recently) east of India (Cordex *et al.* 2004), substantiating a linguistic correlation with the TB speakers of Southeast Asia.

10. Applications of studying Y-chromosomal Diversity:

10.1. Anthropology/ Molecular Evolution

The field of molecular anthropology is predicated on the concept that the genes of modern populations encode aspects of human history. By studying the degree of genetic molecular variation in modern organisms, one can, in principle, understand past events. The Y chromosome is uniquely suited to such studies. The unique nature of Y-chromosome diversification provides an elegant record of human population histories allowing researchers to reconstruct a global picture, emblematic of modern human origins, affinity, differentiation, and demographic history.

10.2. Genealogical Investigations:

Y chromosome variation studies are very useful in genealogical reconstruction (verifying membership in a particular family's ancestry). Since most of the Y chromosome has the special property of not recombining during meiosis, no shuffling of DNA from different ancestors occurs. As a consequence, any Y chromosome accumulates all the mutations that have occurred during its lineal life span and thus preserves the paternal genetic legacy that has been transmitted from father to son over the generations. The discovery of numerous Y chromosome polymorphisms has allowed us to deduce a reliable genealogy composed of numerous distinctive lineages. This concept is analogous to the genealogical relationships maintained by the traditional transmission of surnames in some cultures, although the gene tree approach provides access to a prehistorically deeper set of paternal relationships. In many societies, surnames, like Y chromosomes, are patrilinearly

inherited. This has led to considerable interest among amateur genealogists in whether branches of a family tree can be reliably connected using DNA evidence (Jobling 2001).

10.3. Forensic work:

Due to the recent technological advancement in the biological sciences and the introduction of commercially available multiplex different genotyping kits, Y chromosome is used for identifying male genetic components in material from crime scenes, especially from sexual assault and rape cases. They are also applied to paternity testing, especially in deficiency cases with a deceased alleged father.