Mitochondria and Human Evolution

BY

MAX INGMAN
Dissertation presented at Uppsala University to be publicly examined in Rudbeckalen, Rudbeck Laboratory, Uppsala on 11th October 2003 at 09:15 for the Degree of Doctor of Philosophy (Faculty of Medicine) in Medical Genetics.

ABSTRACT


Mitochondrial DNA (mtDNA) has been a potent tool in studies of the evolution of modern humans, human migrations and the dynamics of human populations over time. The popularity of this cytoplasmic genome has largely been due to its clonal inheritance (in Man) allowing the tracing of a direct genetic line. In addition, a comparatively high rate of nucleotide substitution facilitates phylogenetic resolution among relatively closely related individuals of the same species.

In this thesis, a statistically supported phylogeny based on complete mitochondrial genome sequences is presented which, for the first time, unambiguously places the root of modern human mitochondrial lineages in Africa in the last 200 thousand years. This conclusion provides strong support for the “recent African origin” hypothesis. Also, the complete genome data underline the problematic nature of traditional approaches to analyses of mitochondrial phylogenies.

The dispersal of anatomically modern humans from the African continent is examined through single nucleotide polymorphism (SNP) and sequence data. These data imply an expansion from Africa about 57 thousand years ago and a subsequent population dispersal into Asia. The dispersal coincides with a major population division that may be the result of multiple migratory routes to East Asia.

Also investigated is the question of a common origin for the indigenous peoples of Australia and New Guinea. Previous studies have been equivocal on this question with some presenting evidence for a common genetic origin and other proposing separate histories. Our data reveal an ancient genetic link between Australian Aborigines and the peoples of the New Guinea highlands.

Key words: mitochondria, hominidae, population genetics, human evolution.
MAIN REFERENCES

This thesis is based on the following papers:


These papers are reproduced in the appendices at the back of this thesis. Reprints were made with permission from the publishers.
## ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>D-loop</td>
<td>displacement loop (mitochondrial control region)</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase (thousand base pairs)</td>
</tr>
<tr>
<td>kya</td>
<td>thousand years ago</td>
</tr>
<tr>
<td>Mb</td>
<td>megabase (million base pairs)</td>
</tr>
<tr>
<td>MRCA</td>
<td>most recent common ancestor</td>
</tr>
<tr>
<td>mtDNA</td>
<td>mitochondrial DNA</td>
</tr>
<tr>
<td>Mya</td>
<td>million years ago</td>
</tr>
<tr>
<td>NRY</td>
<td>non-recombining region of the Y chromosome</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>RFLP</td>
<td>restriction fragment length polymorphism</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>TMRCA</td>
<td>time to most recent common ancestor</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>UEP</td>
<td>unique event polymorphism</td>
</tr>
</tbody>
</table>
# CONTENTS

## ABSTRACT

## MAIN REFERENCES

## ABBREVIATIONS

## CONTENTS

## INTRODUCTION

### Background
- Anthropology
- Early molecular studies

## Molecular Population Genetics
- Neutral theory of molecular evolution
- Sources of genetic variation
- Estimation of genetic distance
- Nucleotide substitution rate
- Phylogenetic reconstruction
- Population size

## MITOCHONDRIA AND HUMAN EVOLUTION

### Background
- What are mitochondria?
- Why use mitochondria?
- Mitochondrial evolution
- Recombination in mitochondria
- Natural selection

### Inferences on Human Evolution
- RFLP
- D-loop sequencing
- What about the Y?
- Out of Africa
THESIS WORK

   Paper I – African origin and “mitochondrial Eve”  24
   Paper II – Heteroplasmy and selection in mtDNA  26
   Paper III – Evolutionary history of Aboriginal Australians  28
   Paper IV – Ancient human migrations  30

Concluding Remarks  32

ACKNOWLEDGEMENTS  34

REFERENCES  35

APPENDICES

   Appendix 1 – Paper I
   Appendix 2 – Paper II
   Appendix 3 – Paper III
   Appendix 4 – Paper IV
It has been confidently asserted that man’s origin can never be known: but ignorance more frequently begets confidence than does knowledge: it is those who know little, and not those who know much, who so positively assert that this problem or that problem will never be solved by science.

Charles Darwin
The Descent of Man
1871
INTRODUCTION

Background

Anthropology

Anthropology is defined as “the study of the human race, its culture, society and physical development”. Both cultural and physical anthropology have provided a wide range of starting points for genetic studies of human evolution and migration. These include not only specific data on the cultural and physical history of particular human populations, but also help in understanding general concepts relating to the evolution and spread humans and proto-humans. For example, the biological and behavioural advances that allowed *Homo ergaster* to become the first hominid species to colonise all of Africa and ultimately exit Africa altogether. Figure 1 represents two possible scenarios for the evolution of modern humans based on fossil finds. Both parts of this figure show an origin of *Homo sapiens* in Africa in the last 200 thousand years and subsequent dispersal across the world. An alternative to this view is that modern humans evolved in parallel from *Homo erectus* in multiple locations in both Africa and Asia. Proponents of this “multiregional” view maintain that the similarities seen between extant modern humans from different regions are the result of admixture between populations.

![Figure 1](image_url)

**Figure 1.** Geographical and temporal distribution of hominid populations, based on fossil finds, using different taxonomic schemes. Scenario a) reflects the view that both Neanderthals and modern humans derived from a widespread ancestral species called *Homo heidelbergensis*. Scenario b) represents the alternative stance that Neanderthals have deep roots in Europe and might extend back over 400 kya and that the ancestor of modern humans was a separate African species called *Homo rhodesiensis* (redrawn after Stringer 2003).
In recent years, paleoanthropologists have discovered human fossils antedating 4 million years ago that show the ape origins of the human family ever more clearly and have discovered archaeological sites that imply a complex history of branching events within the genus *Homo*. Of key importance to genetic data, is the discovery of a fossil of the oldest known hominid (australopithecines) in Africa in 1924 (Dart 1925). Since then, more of these fossils have been uncovered but knowledge about the australopithecines remains sparse. Of particular interest is that they lived between 5 and 1.2 million years ago and had a morphology that was intermediate between apes and humans. These fossils place the timeframe for the split between apes (chimpanzees) and humans at 5 to 7 million years ago (Klein 1999).

Emerging theories on the route and timeframe for a modern human exodus from Africa are based on physical evidence such as the discoveries of artefacts which imply the human colonisation of Ethiopia in East Africa at least 125 kya (Walter *et al* 2000). These coastal people probably took advantage of marine food resources, which suggests a major change in the human ability to adapt to local environments. This would also have equipped early *Homo sapiens* with the necessary technology to enable them to migrate from Africa following a coastal route, and eventually adapt their tools and society to enable them to exploit new environments (Stringer 2000). This model has a particular impact on early migration to Australia. The discovery of burial sites at Lake Mungo, variously estimated to be 40-50 kya (Bowler *et al* 2003) and 62 kya (Thorne *et al* 1999) corroborates the assertion that Australia has been colonised for at least 50 thousand years (Roberts *et al* 1990). Since there were no land bridges from South East Asia to Australia, colonisation must have required the use of boats to cross stretches of open sea of up to 100 km wide, implying the expertise of a coastal community.

**Early molecular studies**

From the time of Charles Darwin, it has been the dream of many biologists to reconstruct the evolutionary history of all organisms on Earth and express it in the form of a phylogenetic tree. Lacking a complete fossil record, many investigators have employed the methods of comparative morphology and comparative physiology. However, the evolutionary change of morphological and physiological characters is complex and somewhat subjective so phylogenetic trees reconstructed from these data have often been controversial. Advances in molecular biology changed this situation and have allowed scientists to study the genetic variation between species.

Prior to the advent of methods for rapid DNA sequencing in 1977 (Maxam and Gilbert 1977; Sanger *et al* 1977), most studies of molecular evolution were conducted using amino acid sequence data. Although amino acid sequencing was time consuming and error prone, some important principles of molecular evolution, such as gene duplication and the molecular clock, were discovered. Currently, most molecular studies employ DNA sequencing and, if an amino acid sequence is required, it can be deduced from the DNA sequence.

The first attempt at reconstructing human evolution on the basis of genetic data from living populations was undertaken in 1964 (Cavalli-Sforza and Edwards 1964). This
study presented genetic distances between pairs of selected populations for as many genes as knowledge at the time permitted. During the 1970s more genes were added and HLA data were used to confirm the results of the first investigation (Piazza et al 1975).

The first DNA polymorphisms studied in humans for evolutionary purposes were from mitochondrial DNA (mtDNA). In this study, Brown (1980) proposed the existence of population specific patterns of restriction enzyme cleavage and, using an estimated mitochondrial substitution rate (Brown et al 1979), suggested that the observed amount of sequence heterogeneity could have been generated from a small mitochondrially monomorphic population that existed about 180 thousand years ago.

A groundbreaking study in 1987 extended on this proposal. Cann et al (1987) presented genetic evidence that the last shared ancestor of all contemporary humans existed in Africa no more than 200 thousand years ago. This lead to the suggestion that a woman existed in Africa about 200 thousand years ago who was the ancestor of all extant humans (“mitochondrial Eve”). This is a misnomer as there is no evidence that the human population went through a bottleneck that reduced its number to one (or a few) women. This study has been criticised for a lack of statistical support for the tree topology and for an incorrectly applied statistical analysis (Templeton 1992; Nei 1992). Remedial analyses produced many tree topologies, some with an African root and some rooted elsewhere (Hedges et al 1992). Lacking sufficiently strong empirical data, it is impossible to confidently place the root of modern human mtDNA lineages in sub-Saharan Africa.

Molecular Population Genetics

Neutral theory of molecular evolution

Up until the 1960s, it was believed that natural selection was the primary driving force of molecular evolution.

*The consensus is that completely neutral genes or alleles must be very rare if they exist at all. To an evolutionary biologist, it therefore seems highly improbable that proteins, supposedly fully determined by genes, should have nonfunctional parts, that dormant genes should exist over periods of generations, or that molecules should change in a regular but nonadaptive way ... [natural selection] is the composer of the genetic message and DNA, RNA, enzymes, and other molecules in the system are successively its messengers.*

(Simpson 1965)

There were two schools of thought at this time. The “Classical School” maintained that polymorphism was relatively rare and that evolution consisted mainly of mutation. Mutations were thought to be mostly deleterious and quickly lost through purifying selection but, if advantageous, the effect of positive selection would lead to fixation of the new allele in the population. The “Balance School” was lead by a few people such as Dobzhansky who contested that polymorphisms were quite common and important.
in evolution. They believed and showed that morphological and chromosomal polymorphisms were retained because of heterozygous advantage or other forms of balancing selection. When scientists began to measure directly, they were surprised to find large amounts of genetic variation in natural populations, which was at odds with the expectation under the selectionist (or classical) view.

In 1968, Motoo Kimura suggested that most of the polymorphisms observed at the molecular level are selectively neutral, so that their dynamics in a population are determined by genetic drift (Kimura 1968). This would apply to most nucleotide or amino acid substitutions that occur within a molecule within the course of evolution. The neutral theory models the fate of mutations that are so nearly selectively neutral that their destiny is determined mainly through random genetic drift. If a population of size \( N \) contains a neutral mutation with allele frequency \( p_0 \), then the probability that the mutation reaches fixation in that population equals \( p_0 \). A new mutation will originally exist as a single copy so the initial allele frequency will be \( p_0 = 1/(2N) \). This is also the probability of eventual fixation of the allele. Therefore, where the population size is small, the new allele will have a greater chance of reaching fixation. The rate at which new mutations are fixed in the population equals \( \mu \), where \( \mu \) is the \textit{neutral mutation rate}. The rate of fixation depends only on mutation rate and not on population size as it is determined by the product of the probability of fixation of a new neutral mutation and the average number of new mutations in each generation \((2N\mu)\),

\[
\frac{1}{(2N)} \times (2N\mu) = \mu.
\]

If the rate of fixation is \( \mu \) per unit time, then the average length of time between substitutions will be \( 1/\mu \). Therefore, a higher mutation rate will decrease the time interval between fixations. The mutation rate of mammalian mitochondrial DNA is about 5 to 10 times higher than in the nuclear genome (Brown \textit{et al} 1979; Brown \textit{et al} 1982) and consequently, the time between fixations of new neutral mutations will be shorter.

For each new neutral mutation that is fixed, the average time taken to reach fixation is \( 4N_e \) generations (where \( N_e \) is the effective population size). This means that in a small population, not only will there be a higher chance of fixation, but it will also occur more quickly than in a larger population. Since mitochondria are haploid and are uniparentally inherited, their population size is a quarter of that of the autosomes so the chance of fixation of a new neutral mutation will not only be higher but it will proceed to fixation more quickly. If the probability of fixation of a neutral allele by drift, \( 1/(2N) \), is greater than selection pressure, then drift will outweigh selection. Therefore, regions of the genome that are less constrained by selection should have a higher rate of sequence evolution. In most cases this is true as synonymous substitutions generally occur in a population much more frequently than nonsynonymous substitutions. This means that the majority of sequence evolution is through mutational changes that have little or no effect on phenotype or fitness.
If the neutral theory holds, then it is expected that amino acid evolution (substitution) should be linear with time. That is, if $\mu$ is constant across species and most protein evolution is neutral, then we would expect the rate of molecular evolution to be approximately the same in all lineages. This is known as the molecular clock (Zuckerkandl and Pauling 1962) and is illustrated in Figure 2.

![Figure 2](image)

**Figure 2.** Relationship between the estimated number of amino acid substitutions in the $\alpha$-globin gene of various species relative to humans and the time humans diverged from each species (data from Kimura 1983).

However, the rate of mutation in the genome of a particular organism is not uniform across all genes. The neutral theory predicts that the rate of molecular evolution is dependant on the neutral mutation rate which can vary widely depending on selective pressure.

The hypothesis of neutral evolution can be tested by comparing observed and expected amounts of genetic variation. If the difference between these two values is significantly large, then the action of selection can be assumed. A commonly used test for deviation from selective neutrality was proposed by Tajima in order to examine the relationship between two measures of genetic variation (Tajima 1989a). Nucleotide diversity per sequence ($\pi$) and the number of segregating sites per sequence ($K$) can be calculated from the dataset. Either of these variables can be used to estimate $\theta$ (nucleotide polymorphism ($= 4N\mu$)) and the difference between them divided by the square root of their variance gives Tajima’s $D$ statistic,

$$D = \frac{\pi - K / a}{\sqrt{V(\pi - K / a)}}$$

where $a = 1 + 2^{-1} + 3^{-1} + ... + (m-1)^{-1}$ in a sample of $m$ sequences.
Since $K$ ignores the frequency of variants, it is strongly affected by the existence of low frequency deleterious alleles. In contrast, $\pi$ is dependent on the frequencies of variants and is therefore not particularly affected by the presence of deleterious alleles since they have only a small overall contribution compared to higher frequency neutral alleles. If the alleles in the sample are selectively neutral and the population is in mutation-drift equilibrium, then these two estimations will not be significantly different and $D$ will be close to zero. Under purifying selection, $\theta$ estimated from $K$ will be inflated whereas $\theta$ estimated from $\pi$ will not be affected and a negative $D$ value will result. If the region is under positive selection, the converse will be evident. However, if the population passes through a bottleneck, $D$ may become significantly negative as nucleotide diversity is more strongly affected relative to the number of segregating sites (Tajima 1989b). This is because many variants with low frequency might still be observed in the population. Also, if a population experiences rapid expansion, a significantly negative $D$ value can result due to an increase in the number of segregating sites with little increase in nucleotide diversity. In addition, significantly negative $D$ values can result from a range expansion where there is a high rate of emigration between demes (Ray *et al* 2003). Although this test should not be used indiscriminately, it can be useful for inferring something about a population’s history.

**Sources of genetic variation**

The processes that produce genetic variation in mitochondrial DNA (mtDNA) are somewhat more limited than those that affect nuclear DNA. For example, recombination mixes paternal and maternal nuclear DNA into the chromosomes of their progeny. All nuclear chromosomes experience this effect to some degree with the exception of the majority of the Y chromosome which is inherited only from the father†. Mitochondria have also been shown to be inherited uniparentally but from the maternal line (Giles *et al* 1980).

Mutation is the single most important factor in producing genetic variation in mtDNA. There are three basic types of mutational changes: *substitution* of one nucleotide for another nucleotide, *deletion* of one or more nucleotides, and *insertion* of one or more nucleotides. If insertions or deletions occur in a protein-coding region, they are referred to as *frameshift mutations* since they alter the reading frame downstream of the mutation (unless 3x nucleotides are inserted/deleted). Nucleotide substitution is the primary source of mitochondrial polymorphism that is employed in studying genetic relationships between individuals and can be subdivided into two classes: *transitions* and *transversions*. A transition is the substitution of a purine (adenine or guanine) for another purine or a pyrimidine (thymine or cytosine) for another pyrimidine. Substitution of a purine for a pyrimidine or vice versa is referred to as a transversion. If the substitution occurs in a protein-coding region then it can be either *nonsynonymous* or *synonymous*. Nonsynonymous or replacement substitutions are those that result in an amino acid replacement at that codon (a triplet of three bases

---

† the non-recombining region of the human Y chromosome (NRY) comprises 95% of the chromosome and does not undergo sexual recombination.
that codes for an amino acid). Synonymous or silent substitutions do not result in an amino acid change. This is due to codon degeneracy which relates to the fact that there are 61 possible sense codons (a codon that doesn’t result in protein translation termination) that code for only 20 amino acids. This means that different combinations of 3 nucleotides can code for the same amino acid so a nucleotide substitution will not necessarily result in an amino acid substitution. In addition, there are nonsense mutations which are nonsynonymous mutations that result in a stop codon, terminating protein translation. A comparison of the number of synonymous substitutions per site \( (k_s) \) with the number of nonsynonymous substitutions per site \( (k_a) \) conveys information regarding the type of selection that could be acting on a sequence. In the presence of selection that is preserving the integrity of the gene product through functional constraints (purifying selection), \( k_s \) is expected to be significantly larger than \( k_a \). Often this is represented as a ratio of \( k_a/k_s \) where a value of less than 1 indicates the action of purifying selection.

**Estimation of genetic distance**

Although nucleotide substitutions may accumulate in a genome at a regular rate, for most organisms the rate is too slow to measure within an individual. For this reason, nucleotide substitutions are usually detected by comparing sequences that have derived from a common ancestral sequence. The descendant sequences gradually diverge by nucleotide substitution. Analysis of the number of differences that have accrued between the sequences is the basis for the reconstruction of phylogenetic trees and allows the rate of substitution to be calculated. However, estimation of the number of nucleotide substitutions between a pair of sequences is rarely a straightforward task. The simplest measure of sequence divergence is referred to as the \( p \) distance and is calculated by dividing the number of observed substitutions between a pair of aligned sequences \( (\hat{K}) \) by the number of nucleotide sites examined (the sequence length). This measure gives the number of nucleotide substitutions per site assuming that every site in the sequence is equally likely to undergo substitution and be replaced by any other nucleotide. As discussed earlier, there are two classes of substitution: transitions and transversions. In most DNA sequences, transitions occur much more frequently than transversions so it becomes necessary to introduce an extra parameter to account for this. The ratio of transitions to transversions is about 0.5 to 2 in nuclear genes but can be as high as 25 in mitochondrial DNA (Excoffier and Yang 1999). In addition, if the value for \( p \) is very high, there is a likelihood that the same site may have undergone substitution more than once (multiple substitution). If this occurs in a single lineage it is referred to as a back mutation since the most likely outcome is that the site will be returned to its original state (eg. A→G then G→A at the same site). Another confounding factor coupled to a high \( p \) value is parallel substitution where the same site undergoes the same nucleotide substitution in multiple lineages independently. In either case, two substitutions have occurred but neither will be observed in a pairwise comparison. Both of these situations will lead to underestimation of the genetic distance between the sequences and will affect the calculation of the time at which they diverged. In a sample of 53 mitochondrial control region sequences, it was found
that some sites had undergone parallel substitution in up to ten separate lineages (Ingman and Gyllensten 2001). This means that at some loci, a large percentage of sequences share polymorphisms by chance rather than by descent.

The problem of multiple substitutions is illustrated in Figure 3. While the substitution rate has remained relatively constant, the accumulation of substitutional changes over time between these species is not linear. In fact, the rate of accumulation of pairwise synonymous substitutions ranges from 17 per million years in species that diverged 2 million years ago to 4.5 per million years in species that diverged 20 million years ago (Janecek et al 1996). This is due to a saturation of mutations, making multiple substitutions more common. In agreement with the molecular clock hypothesis represented in Figure 2, the rate of accumulation of nonsynonymous substitutions (and therefore amino acid replacements) remained fairly constant.

![Figure 3](image.png)

Figure 3. Total nucleotide substitutions between pairs of mitochondrial cytochrome c oxidase subunit II gene (684 bp) from various bovid taxa versus time since divergence (from Janecek et al 1996).

Of course, this is an extreme example and, because there is considerably less time depth, the problem is not so pronounced amongst human mitochondrial DNAs. Nevertheless, due to these complications, it is necessary to use a mathematical model to estimate the number of nucleotide substitutions between pairs of sequences.
Figure 4. Estimates of the number of nucleotide substitutions obtained by different distance measures (Tamura 1992; Kimura 1980; Jukes and Cantor 1969) when actual nucleotide substitution follows the Tamura-Nei model (Tamura and Nei 1993). The higher the complexity of the model, the more accurately it tracks the actual distribution as the number of nucleotide substitutions increases (from Nei and Kumar 2000).

These models vary in their complexity and their estimates diverge dramatically as $p$ increases. As Figure 4 shows, more complex models perform better when $p$ is large and certainly all models outperform the $p$ distance. More complex models typically allow for extra parameters such as unequal nucleotide frequencies, unequal transitional and transversional substitution rates and transition/transversion ratio. Unfortunately, increasing the complexity of the model also increases the variance that the model produces. This means that the choice of model should be dependent on the dataset to which it is to be applied. Therefore, when $p$ is small a simple model is best and it becomes necessary to employ successively more complex models as $p$ increases.

These calculations of genetic distance assume that the rate of nucleotide substitution is uniform across all sites in the sequence. This is rarely the case and the rate is known to vary widely between coding and non-coding regions as well as at different codon positions within coding regions. Adding to this increasingly complicated picture is the presence of “hotspots” and “coldspots” for mutation. These are sites that are undergoing substitution at a rate that is much higher or much lower than the surrounding sites. Since the hotspot sites have a much higher mutation rate, it is more likely that they will have undergone a parallel or back mutation. The substitution rates in the human mitochondrial D-loop are known to be highly heterogenous (Maddison et
al 1992; Tamura and Nei 1993; Wakeley 1993) and can vary by as much as a factor of 6 between adjacent sites (Meyer et al 1999). Some of these substitution models can accept an extra parameter, gamma, which allows each site to be binned into a rate category for more accurate estimation of genetic distance in regions with extreme rate heterogeneity.

**Nucleotide substitution rate**

Initially, the neutral theory of molecular evolution was based on observations of the rate of amino acid replacement in proteins. The rate of substitution in nucleotide sequences can be analysed in much the same manner as amino acid sequences and is a basic quantity in the study of molecular evolution. Comparisons of substitution rates between genes and among different regions of a gene can assist in understanding the mechanism of nucleotide substitution in evolution. Nucleotide substitution rate can also be used to date evolutionary events such as species divergence or the time since members of the same species shared a common ancestor. In these instances, the assumption of a molecular clock is of utmost importance.

The rate of nucleotide substitution can be defined as the number of substitutions per site per year. Given the genetic distance \( K \) between two homologous DNA sequences and the time that has elapsed since divergence \( T \), the rate of divergence \( r \) of the two sequences can be easily calculated,

\[
r = \frac{K}{2T}.
\]

The reason the genetic distance is halved is because nucleotide substitutions have occurred on both lineages independently since their divergence and it is assumed that the substitutions that have occurred on each lineage will be half this value. Even though genetic divergence typically precedes population divergence, the divergence time for the two sequences is often assumed to be the same as the time of divergence of the two species inferred from palaeontological data.

When the genetic distance between sequences is small, long sequences are required in order to minimise stochastic error. If the time since divergence of the two sequences is small, then a slight error in estimation of \( T \) can lead to a substantial error in the rate of substitution. Due to this, it is necessary to estimate substitution rates with sequences that are not closely related. However, as discussed previously, it is difficult to estimate genetic distance when the divergence occurred a very long time ago due to the effect of multiple substitutions. Deciding on the optimal species for this comparison becomes of high importance but the choice is often made on the basis of what palaeontological data are available. For example, the closest living relatives to humans are chimpanzees (Saitou and Nei 1986) and paleoanthropologists have dated their divergence on the basis of a restricted number of australopithecine fossils at 5 to 7 million years ago (Klein 1999). It is this estimate, along with the average genetic distance between humans and a chimpanzee sequence that is commonly used to estimate substitution rates for various human genes. Of course, the substitution rate will vary depending on the time that is used for the species split and depending on the problems in estimating
genetic divergence described in the preceding section. In addition, it is necessary to assume that both species have the same neutral mutation rate or the substitution rate estimate will fall midway between the rates for the two species. This situation can arise if the two species have, for example, different generation times.

Another confounding factor in the estimation of nucleotide substitution rates is rate heterogeneity. This refers to the situation where the rate of substitution is not uniform along the length of the sequence. Failure to take rate variation among sites into account can lead to a severe underestimation of the amount of sequence divergence. However, it is possible to roughly verify estimates of substitution rates directly using data from mutations in pedigrees. This has been calculated for mitochondrial DNA and the results affirm that the substitution rate that has been derived from phylogenetic studies is at least approximately correct (Parsons et al 1997; Sigurðardóttir et al 2000).

**Phylogenetic reconstruction**

Molecular phylogenetics is defined as the study of the evolutionary relationships among organisms or genes by a combination of molecular biology and statistical techniques (Li 1997). A common representation of this evolutionary relationship is the reconstruction of a phylogenetic tree. A phylogenetic tree is similar to a family tree in that it displays evolutionary relationships with ancestral lines represented as the branches of a tree. In the simple tree presented below in Figure 5, E is the common ancestor for all of the other taxa. At \( t_1 + t_2 \) years ago, the lineage E diverged into two descendant lineages. The lineage on the left eventually evolved into A, whereas the lineage on the right evolved into D after \( t_1 \) years before then splitting again to eventually evolve into both B and C.

![Figure 5. A simple phylogenetic tree where A, B and C are extant taxa and D and E are ancestral.](image)

Reconstruction of phylogenetic trees using statistics developed in parallel in the fields of numerical taxonomy for morphological characteristics and in population genetics during the 1960s. Although some of the early methods have become antiquated, a number of the concepts are still applied in modern tree building techniques. An important distinction to make is the difference between the two types of tree. A *species tree* is a phylogenetic tree that represents the evolutionary pathways of a group
of species whereas a gene tree is reconstructed from a gene from each species or a gene each from several members of the same species. One of the main differences between the two is illustrated below in Figure 6. Since genetic divergence will normally precede population divergence, dates based on a gene tree will usually predate than those based on a species tree by an amount that reflects the level of polymorphism in the ancestral species (Takahata and Nei 1985). In the presence of polymorphism, it becomes necessary to reconstruct a tree based on sequence data from multiple independent loci (Saitou and Nei 1986). When more than two genes are sampled from each population, it is possible to make a correction for the effect of ancestral polymorphism. The number of nucleotide substitutions (net substitutions), $d$, that have occurred since population splitting can be estimated by subtracting the number of within-population differences (Nei and Li 1979).

![Figure 6. Diagram showing the time of gene splitting (gs) is usually earlier than the time of population splitting (ps) when polymorphism exists (from Takahata and Nei 1985).](image)

Secondly, the pathway of genetic divergence is not always the same as that for species divergence. That is, the topology of the two trees can be considerably different. In the three examples in Figure 7, all of the species trees are the same but the gene trees show different topologies (A and B versus C) and different genetic time depth (A versus B and C). This is because nucleotide and amino acid substitutions occur stochastically so to avoid topological errors, it is necessary to examine a large number of nucleotides or amino acids. Ideally, a gene tree should contain sequence data from several independent loci since a tree reconstructed from a single gene may still differ from the species tree even if a large number of nucleotides are used (Saitou and Nei 1986).
Figure 7. Three possible relationships between species trees and gene trees for the case of three species in the presence of polymorphism. The times of the first and second splitting are $t_0$ and $t_1$, respectively (from Nei 1987).

Because many amino acid substitutions won’t cause a phenotypic change, molecular data offer more information for evolutionary studies than morphological or physiological data and, since many nucleotide substitutions won’t cause an amino acid change, nucleotide data contain more information than protein data. Since nucleotide substitutions occur randomly, to minimise stochastic errors long DNA sequences are required.

Part of the problem of inferring the correct branching topology is that the number of possible bifurcating trees increases very rapidly as additional taxa are added. For example, for five taxa there are 105 possible trees and for ten taxa there are 34,459,425. Clearly, computers have made this task easier but even the fastest computer could not examine every topology for a dataset of more than thirty taxa. Because of this, many tree building algorithms try to eliminate classes of trees that are dependent on each other based on the observed data. There are three main groups of methods that are commonly used for the reconstruction of phylogenetic trees using nucleotide data: distance, maximum parsimony and maximum likelihood methods.

Estimation of branch lengths is relatively easy compared to the problems faced with attempting to reconstruct the topology of the true tree from a dataset of aligned DNA sequences and each of these groups of methods uses a different basic principle to attempt this. Distance methods use estimates of genetic distance between all pairs of sequences in the dataset and the tree is reconstructed using an algorithm based on functional relationships among the distance values. Maximum parsimony is concerned with finding the best tree, which is the tree with the shortest evolutionary pathway to variant sites that are shared by two or more sequences (parsimony informative sites). Maximum likelihood methods calculate the likelihood of the dataset being derived from each possible tree and choses the highest value as the best tree.

Homoplasy is when two or more sequences in a dataset have the same variant at the same site through parallel mutation. When this occurs in a dataset, the reliability of a
reconstructed tree declines (Saitou and Nei 1986). Parsimony methods are more sensitive to homoplasy since they are only utilising shared variant sites and attempting to explain the distribution of these in the simplest possible manner. While both parsimony and likelihood methods will often produce more than one best tree (sometimes thousands with the same tree score), the distance method will produce just one. It should be remembered that this one tree is still just an estimation of the true tree and statistical tests are required to analyse the stability of the topology. Nevertheless, distance methods have become commonly used for phylogenetic reconstruction from nucleotide data, particularly in human population genetics. The most common tree building algorithms that employ distance matrices are UPGMA (Sokal and Michener 1958) and Neighbor-Joining (Saitou and Nei 1987).

*Population size*

In a natural population, population size is not constant over time and it is unlikely that all individuals contribute gametes to the next generation with equal probability. The *effective population size* \( (N_e) \) of an actual population is defined as the number of individuals in a theoretically ideal population having the same magnitude of random genetic drift as the actual population. Wright (1931) first calculated the effective population size by considering the effective degree of inbreeding in various situations and this is the concept that is currently most commonly used. Simply, effective population size is just the number of individuals in a population that can contribute genetically to the next generation at any one time. In general, \( N_e \) is considerably smaller than the actual population size \( (N) \). This is largely due to the incidence of overlapping generations where at any one time, a population will contain individuals in pre-reproductive, reproductive and post-reproductive stages. Another important contributing factor is differential male and female reproductive probability due to polygamy (polyandry or polygyny) or uneven numbers of the sexes. Also, in humans only a male can pass on a Y chromosome and only a female can contribute a mitochondrion so the \( N_e \) for these loci will be the effective population size of males and of females, respectively.

As mentioned above, population size will rarely remain constant but will generally fluctuate over time and therefore the effective population size will also fluctuate. A population *bottleneck* is indicated by a dramatic decrease in population size and results in marked loss of heterozygosity and nucleotide diversity, and increased levels of linkage disequilibrium. This often occurs when a small group of emigrants leaves a population and founds a new subpopulation. Because of the small size, random genetic drift will have a strong influence in this new population and this is referred to as *founder effect*. Bottlenecks followed by founder effects in human populations could contribute to regional differentiation in morphology and the high incidence of an otherwise rare genetic disease in a population.

As noted previously, a bottleneck can be detected by tests of selective neutrality. This is indicated by a significantly negative Tajima’s \( D \) statistic due to a reduction in nucleotide diversity relative to the number of segregating sites. A rapid increase in population size can also be detected in this way due to an increase in the number of
segregating sites with little increase in nucleotide diversity so it is therefore necessary to apply other analyses. An analysis of the distribution of pairwise differences between all sequences in a population sample has become a common way to infer population history. This is referred to as a *mismatch distribution* and is simply a histogram of the frequency of the number of nucleotide differences between all pairs of sequences. In the early 1990s, several investigators examined pairwise comparisons of mitochondrial DNA sequences in stable and in growing populations (Slatkin and Hudson 1991; Rogers and Harpending 1992; Harpending *et al* 1993; Harpending 1994). They found that under a neutral infinite-sites model with no recombination and constant population size, the distribution of the number of differences showed a ragged pattern (Figure 8 top). However, the mismatch distributions for populations that had undergone a sustained period of growth are smooth and have a peak (Figure 8 bottom) and the position of the peak reflects the time of the population growth.

![Figure 8. Distribution of pairwise sequence differences within 6 African populations. The x-axis shows the number of nucleotide differences between pairs of mtDNA HVR1 sequences and the y-axis gives the number of pairs. The 3 top populations are stable while the bottom 3 have experienced a period of population growth (from Watson *et al* 1996).](image)

It should be noted that simulations have shown that unimodal distributions like those represented in Figure 8 (bottom) can also be obtained after a range expansion followed by a high rate of emigration to neighbouring demes (Ray *et al* 2003).
From the pairwise mismatch distributions in Figure 8, stark differences in population history are seen among hunter-gatherer populations like !Kung, Mbuti pygmies and Biaka pygmies, when compared to other African populations that experienced a period of rapid expansion due to a transition to agricultural economies. If the substitution rate is known for the sequences to be compared, then the time to the modal number of differences can easily be calculated. This peak would correspond to the period of highest population growth.

A common method of calculating the time to when this population expansion peak occurred is from an estimate of the parameter Tau (τ). Tau is described as the time that has passed since the population growth or decline measured in units of mutational time,

\[ \tau = 2\mu t \]

where \( \mu \) is the substitution rate (per sequence per generation) and \( t \) is the number of generations (Rogers and Harpending 1992). Tau can be estimated from the dataset and when the substitution rate for the DNA sequence is known, the time since the population growth or decline can be calculated.
MITOCHONDRIA AND HUMAN EVOLUTION

Background

What are mitochondria?
Mitochondria are organelles that exist in the cytoplasm, outside the cell nucleus, and are therefore completely separate from the nuclear DNA. Each cell in the body contains 10 to 100 mitochondrial compartments with each of these housing between 0 and 11 copies of the mitochondrial genome (Cavelier et al 2000). Even though there are a large number of copies per cell, the presence of mitochondrial copies differing by one or more nucleotide substitutions in the same individual (heteroplasmy) is not common. Mitochondria function in the cell as a centre for oxidative phosphorylation (OXPHOS) which is the pathway that produces adenosine triphosphate (ATP), which is used as cellular energy for work and body temperature maintenance.

The mitochondrial genome is circular and consists of approximately 16500 base pairs (bp) of double stranded DNA encoding 13 polypeptides, as well as the 22 transfer RNA (tRNA) genes and 2 ribosomal RNA (rRNA) genes necessary for the transcription and translation of the genome (Figure 9). The control region, often referred to as the D-loop (Displacement loop), is about 1,100 bp long (about 7% of the genome) and contains the origin of replication as well as various control elements. Animal mitochondria also contain several hundred enzymes that are used in metabolic functions but these are encoded by nuclear genes and transported to the mitochondria.

![Figure 9. Schematic overview of the genetic arrangement of the mitochondrial genome. Genes which contribute to the 5 protein complexes are indicated with shades of grey.](image-url)
In mammals, mitochondria are usually inherited from the mother through the egg cytoplasm. In humans, the egg cytoplasm contains a large number of mitochondrial particles whereas the sperm head only carries 5 to 10. During fertilisation, these paternal mitochondria are labelled for degradation through the ubiquitin pathway, resulting in an exclusively maternal inheritance pattern (Sutovsky et al 1999). This uniparental mode of inheritance means that recombination between maternal and paternal mitochondria cannot occur.

**Why use mitochondria?**
The analysis of mitochondrial DNA (mtDNA) has been heavily utilised for studies of human evolution. This is largely due to a handful of appealing features that set mtDNA apart from nuclear DNA. Possibly the most important factor is that, like the Y chromosome, mitochondria are uniparentally inherited. While the lack of apparent recombination means that only the history of women in a population can be investigated, it allows the tracing of a direct genetic line, where all polymorphism is due to mutation. The high substitution rate in the mitochondrial genome compounds this advantage and allows for phylogenetic resolution of more closely related individuals than is possible with nuclear DNA sequences of similar length. If mtDNA is evolving 5 to 10 times faster than nuclear DNA, then 5 to 10 times the amount of nuclear sequence data would be required to give the same amount of phylogenetic data. Also, as the mitochondrial genome is of an immediately finite length (~16,500 bp), a wealth of data from the same locus is available for comparison. In fact, over 10 thousand human mitochondrial hypervariable region sequences (from the mitochondrial D-loop) have been produced to date, many for the purpose of forensic analyses. The D-loop is an important source of information for forensic scientists for many of the same reasons that it appeals to population geneticists, as well as some other salient factors. It can be possible to extract mtDNA from sources when the nuclear DNA is degraded beyond use. This is partly because the circular configuration of the genome makes it more stable in nature and partly because many more copies of the mitochondrial genome exist in each cell. In fact, mtDNA has been extracted, amplified and sequenced from the fossilised remains of Neanderthal bones at least 30,000 years old (Krings et al 1997).

**Mitochondrial evolution**
The substitution rate in the mitochondrial genome is typically 5 to 10 times higher than that of nuclear DNA (Brown et al 1979; Brown et al 1982). The reason for this high rate of substitution is thought to be due to either a high rate of nucleotide misincorporation or low DNA polymerase repair efficiency (Lansman and Clayton 1975). Studies on the reliability of mitochondrial DNA replication have shown that the net average fidelity of the human mitochondrial DNA polymerase (pol γ) is about 1 error per 1-20 million base pairs, a rate that is 1000-fold higher than the error frequency expected for the replication of nuclear DNA (Johnson and Johnson 2001a; Johnson and Johnson 2001b). Also, the high turnover rate of mitochondria in tissues (Rabinowitz and Swift 1970) would provide more rounds of replication during which
errors could be generated. Compounding these factors is an enhanced chance of fixation of new mutations due to low functional constraints on mitochondrial products (Brown et al 1979). In addition to the high mutation rate in the mitochondrial genome, the control region of the genome has been estimated to evolve at a rate of 5 times to nearly 20 times faster (Sigurðardóttir et al 2000) than that of the coding region (the other 93% of the genome).

Recombination in mitochondria
As early as 1980, Giles and colleagues demonstrated maternal inheritance in human mitochondrial DNA (Giles et al 1980). This was confirmed in intraspecific mouse crosses, while in interspecific crosses paternal contribution was observed (Gyllensten et al 1991). However, on the basis of phylogenetic analysis (Eyre-Walker et al 1999) and linkage disequilibrium (LD) studies (Awadalla et al 1999), evidence of recombination has been claimed among paternal and maternal human mitochondrial DNA. The results of these studies have been heavily criticised, particularly the latter (Kivisild and Villems 2000; Jorde and Bamshad 2000; Kumar et al 2000; Parsons and Irwin 2000). The comments centred on the small number of sites that had been analysed, a biased dataset, as well as the use of the LD measure, $r^2$, that is over sensitive to allele frequency variation. Reanalyses of the dataset using the standard LD measure $D'$ show that there is no association of linkage disequilibrium with physical distance. Regardless, Awadalla and colleagues have responded to the criticisms and left the issue of recombination an open question (Awadalla et al 2000). They concede, however, that if recombination does occur in human mitochondria, it is infrequent.

Natural selection
A key assumption in estimating evolutionary distances between mitochondrial lineages is that observed pattern of polymorphism is the result of random genetic drift. This implies that natural selection does not play a part in shaping regional mtDNA variation. However, there are several examples of non-neutrality of mtDNA mutations such as cytoplasmic male sterility in plants (reviewed in Budar et al 2003) and in humans, a class of muscle disease known as mitochondrial myopathy (reviewed in Oldfors and Tulinius 2003). Effects of natural selection in mitochondrial coding regions are also evident in the discordance between the numbers of synonymous and nonsynonymous substitutions (Ingman and Gyllensten 2001). Also, when mitochondrial lineages are subdivided into groupings that are thought to reflect their historical habitat, a larger proportion of nonsynonymous substitutions are seen among the “arctic” lineages relative to the “tropical” and “temperate” lineages in the ATPase 6 gene (Mishmar et al 2003). In fact, the “arctic” populations show a two-fold increase in $k_a/k_s$ ratio (the number of nonsynonymous substitutions per nonsynonymous site divided by the number of synonymous substitutions per synonymous site) averaged over all 13 mitochondrial protein coding genes compared to those from “temperate” locations which in turn show double the $k_a/k_s$ ratio of the “tropical” populations of Africa (Ingman and Gyllensten, in preparation). This could be due to changing
environmental pressures on mitochondrial lineages as humans moved from the African climate to exploit new habitats.

*There can, however, be no doubt that changed conditions induce an almost indefinite amount of fluctuating variability, by which the whole organism is rendered in some degree plastic.*

(Darwin 1871)

This may be especially important in cold environments as mitochondrial oxidative phosphorylation (OXPHOS) employs dietary calories to produce ATP, which can be utilised to generate heat for body temperature maintenance. Mitochondrial lineages may therefore vary functionally depending on their long-term environment and the rate of evolution of lineages from different habitats could be quite different, invalidating the assumption of a molecular clock. While this does not appear to be a major problem since the numbers of non-synonymous substitutions are relatively small, these factors should be kept in mind when interpreting mitochondrial population genetics data.

**Inferences on human evolution**

As discussed above, mitochondrial DNA has been an important source of phylogenetic information for studies of human evolution, migration and population differentiation. To date, few studies have presented phylogenetic analyses based on the complete mitochondrial genome but rather have focussed on particular sites that are thought to be variable, or the rapidly evolving mitochondrial control region (D-loop).

**RFLP**

One of the human population genetics studies that stands out above all others employed mtDNA genetic variation to study the evolutionary relationships between individuals of diverse origins. In 1987, a group of researchers at the University of California campus at Berkley presented data that indicated a recent African origin of modern humans (Cann *et al* 1987). This landmark study was based on restriction fragment length polymorphism (RFLP) variation among the mtDNA of 147 humans from five populations. In this study, 12 restriction enzymes were used to cut the mtDNA of each of these individuals at an average of 370 positions. These restriction enzymes recognise a particular DNA sequence and cleave the strand at a specific site within that sequence. The length of the fragments can then be estimated so the position of the site and the nucleotide sequence at the restriction site can be inferred. Using this technique, Cann and colleagues assayed a total of about 9% of the mtDNA of each individual and identified 195 polymorphic sites. The 147 individuals included in the study yielded 133 distinct mitochondrial haplotypes which were assembled into a phylogenetic tree using the maximum parsimony method. The oldest group (*clade*) on the tree was comprised of mtDNA types derived from African people and, given the number of differences between the two most divergent mtDNA types, the common ancestor of all the haplotypes was estimated to have existed (in Africa) between 140 and 280 thousand years ago.
The RFLP technique continued to be popular through much of the 1990s but, although some researchers still employ this method, improvements in rapid sequencing technology it has made it somewhat antiquated.

**D-loop sequencing**

The mitochondrial D-loop comprises about 1,100 bp which make up less than 7% of the complete mitochondrial genome. The D-loop contains two hypervariable regions (HVR1 and HVR2) which account for much of the high substitution rate in this sequence. Sequencing the D-loop for use in molecular population genetics has been popular due to the high degree of polymorphism that can be found in this relatively short sequence. The advantage of copious amounts of published data from the human D-loop continues to make it an attractive source of phylogenetic information. Many of the methods used for analysing sequence data for phylogenetic inference have been specifically developed for dealing with the particular complications specific to the D-loop. The main problem stems from extensive rate heterogeneity that results from an extreme rate of substitution at some sites (Maddison et al. 1992; Wakeley 1993). The consequence of this is a high frequency of multiple substitutions and, as a result, specific algorithms for estimating genetic distance have appeared in an attempt to take this into account (e.g., Tamura and Nei 1993). However, the pattern of polymorphism in this region is still not well understood and studies continue to be plagued with statistical problems due to the high level of homoplasy in the datasets. Nevertheless, information gleaned from D-loop sequences has added significantly to our level of understanding of human evolution and the genetic relationships among populations.

**What about the Y?**

Like the mitochondrial genome, the non-recombining region of the Y chromosome (NRY) is uniparentally inherited and therefore has the advantage of following the genetic line of males in a population without recombination. Its recent origin, coupled with high geographic differentiation (Jobling and Tyler-Smith 1995), make the Y chromosome an important tool in the study of modern human migrations. Since 1995, Y chromosome markers have been employed increasingly in studies of human evolution (Hammer 1995; Jobling and Tyler-Smith 1995). Rather than complete sequencing, data are generally gathered from microsatellite loci and SNP sites. Although SNPs on the Y chromosome are rare (Hammer 1995; Jobling and Tyler-Smith 1995) the NRY is about 35 Mb long and with a pairwise difference on average every 3 to 4000 bases (Stumpf and Goldstein 2001), it represents an almost unlimited source of phylogenetic information. In addition, since the Y chromosome is evolving at a much slower rate than mtDNA, the Y mutations are considered to be unique event polymorphisms (UEPs) which reflects the fact that parallel or back mutations are extremely unlikely. In fact, a study based on the analysis of 167 NRY polymorphisms in 1062 individuals produced 116 haplotypes that could be parsimoniously assembled into a single genealogy (Underhill et al. 2000). The human Y chromosome also offers greater geographic variability than mtDNA since women have had a considerably higher migration rate than men due to the predominance of patrilocality in human
populations (Seielstad et al 1998). Perhaps the only key ingredient missing in studies of the Y chromosome is a full characterisation of variability within that locus allowing greater resolution for more closely related lineages. Presumably, that will not be too far in the future.

_Out of Africa_

General descriptions of the global pattern of the movements of modern human peoples are a synthesis of physical and cultural anthropology with molecular data from multiple genetic systems and loci. Also taken into account are geological factors such as the degree of glaciation which can affect migration paths and raise or lower sea levels. This, for example, allows a time frame for the possibility of human migration from Asia to the Americas via the Bering Strait land bridge.

**Figure 10.** The origin and dispersal of modern humans. Evidence from mitochondrial genomes bolsters the hypothesis that the place of origin was sub-Saharan Africa between 100 and 200 thousand years ago and that the dispersal from Africa occurred within the past 100 thousand years. The earliest known fossil and archaeological evidence on each continent, shown on the map, is consistent with this view (redrawn after Hedges 2000).

The simplified account described in Figure 10 shows directionality of supposed modern human movement based on the earliest physical evidence of modern humans in different regions but the lines on the map should not be mistaken for migration pathways. A general (but incomplete) consensus is that anatomically modern humans evolved in Africa 100 to 200 thousand years ago and exited Africa in the last 100
thousand years to the Levant, the Mediterranean region between Turkey and Egypt. From there, they radiated toward East Asia, Europe, Australia and eventually to the Americas. As modern humans advanced across the world, they must have encountered archaic humans that were resident in their new range. A key question is whether there was total replacement of these previous occupants or if admixture between modern and archaic forms has contributed to the contemporary human gene pool. While barriers to fertility are slow to develop, making admixture a biological possibility, cultural and social barriers were likely to be of greater significance (Cavalli-Sforza et al 1996). Fossil and archaeological evidence suggest that the modern human form evolved prior to the evolution of the modern human facility for culture (Klein 1999), which implies that archaic human forms would have had a somewhat lower disposition for complex cultural expression and a lower level of language skills. It has been postulated that the latter would have imparted modern humans with greater fitness and provided the strongest barrier to admixture (Cavalli-Sforza et al 1988).
**THESIS WORK**

**Paper I – African origin and “mitochondrial Eve”**

The two competing hypotheses on the origin of modern humans agree that *Homo erectus* evolved in Africa about 2 million years ago and subsequently colonised much of Asia and Europe. The “recent African origin” hypothesis then states that modern humans later evolved in Africa some 100 to 200 thousand years ago and eventually spread to the rest of the world, systematically replacing *Homo erectus* without genetic mixing. Proponents of “multi-regional” evolution maintain that the transformation from *Homo erectus* to modern humans occurred in parallel in several different locations outside of Africa. The challenge for molecular anthropologists has been to determine whether the common ancestor of all living humans was in Africa 2 million years ago or 200 thousand years ago. Previous work conducted on human origins has estimated a date for the coalescence of all human mitochondrial lineages of about 200 thousand years ago, however these studies contained some irregularities. An important piece of evidence that was missing from the work of Cann *et al* (1987) and those studies that followed, was statistical support for the topology of the phylogenetic trees, particularly in the deep African branches. The amount of data collected during these analyses of partial mitochondrial sequences simply wasn’t sufficient to confidently place the root of all living people in Africa. Compounding this lack of data were the problems created by the high degree of homoplasy in D-loop datasets. Following studies of RFLP data (Maddison 1991) and D-loop sequence data (Vigilant *et al* 1991), Alan Templeton commented that both datasets gave phylogenetic trees with geographically mixed basal clades, thereby invalidating the original rationale for an African origin (Templeton 1992). A reanalysis of the Vigilant dataset using the Neighbor-Joining tree building algorithm revealed that only 6 nodes on the tree were statistically supported and those only defined small clusters of 2 to 6 individuals (Hedges *et al* 1992). Furthermore, the support for the African branches was very low, a problem that continued through the 1990s.

To address this issue, we obtained complete mtDNA sequences from 53 humans of diverse origins and built a tree using the Neighbor-Joining method (Saitou and Nei 1987). As can be seen in Figure 2 (Appendix 1), the tree reconstructed using our mtDNA sequences has a strongly supported basal branching pattern. The three deepest branches lead exclusively to mtDNA sequences derived from individuals from sub-Saharan Africa, providing compelling evidence for a human mtDNA origin in Africa. The date of the most recent common ancestor (MRCA) of all the mtDNA sequences is 171 ± 50 kya, a timeframe that fits with the hypothesis of a recent African origin of modern humans. This date is compatible with a date for the common ancestral human Y chromosome of 188 kya (51 to 411 kya, 95% CI) (Hammer 1995). A study of about 10 kb of non-coding DNA in 69 individuals from a region of low recombination at Xq13.3 presented a date for the MRCA of 535 ± 119 kya (Kaessmann *et al* 1999). Since the population size of the X chromosome is 3 times larger than that of mtDNA

---

1 many researchers now believe that early *Homo erectus* was a separate species, *Homo ergaster*, and *Homo erectus* evolved from *Homo ergaster* after their exodus from Africa.
or the Y chromosome, a date of about three times older is expected. This difference in population size is due to the fact that at each mating there are three X chromosomes but only one mtDNA type and one Y chromosome (and four of each of the autosomes) that can be passed to the progeny. In line with this, the TMRCA for an autosomal locus is found to be about four times older (750 kya (400 to 1,300 kya, 95% CI)) than that for mtDNA or Y chromosomes (Harding et al 1997). In this manner, the common genetic ancestor of all extant human lineages depends on which locus is examined. Clearly, the process of becoming human did not happen overnight but rather was a gradual process of accumulation of genotypic and phenotypic changes in all genetic systems over many thousands of years.

Since we were in possession of an extensive mitochondrial dataset, we were also able to use these sequences to address the issue of recombination among maternal and paternal mitochondria. We examined linkage disequilibrium (LD) among our complete set of mitochondrial genomes using the LD measure $D'$ and found no correlation between $D'$ and nucleotide distance. We concluded that, at least in our dataset, recombination was not a contributing factor to the evolution of mtDNA.

Also, through application of Tajima’s $D$ statistic and analysis of pairwise mismatch distributions, we identified that non-African sequences had experienced a period of population expansion. From an estimate of Tau we were able to date this to 38,500 years ago, a time that coincides with a period of cultural change.
**Paper II – Heteroplasmy and Selection in mtDNA**

In Paper I, we removed the D-loop sequences from the dataset in all phylogenetic analyses. The reason given was that it did not behave in a “clock-like” manner and was therefore unsuitable for phylogenetic inference. This conclusion was drawn because a test of relative rates between gorilla, chimpanzee and human D-loop sequences showed that the overload of multiple substitutions in that region voided the assumption of a molecular clock, whereas the mitochondrial genomes excluding the D-loop were found to evolve at roughly constant rates. In this manuscript, we took this a step further and compared the D-loop to the coding region of the genome in an attempt to determine the extent to which the D-loop was affected by homoplasy resulting from multiple substitutions. Considering the amount of polymorphic sites in the D-loop (30% of the complete genome total) it should be possible to reconstruct an informative tree using this region. Figure 3 (Appendix 2) shows the contrast in resolution between trees reconstructed using the coding region (a) and using the D-loop (b). This difference is particularly pronounced when branches with low statistical support are collapsed (right trees).

Strong evidence exists for mutational “hot spots” in the mitochondrial control region that affect phylogenies reconstructed using these data (Maddison *et al* 1992; Wakeley 1993; Meyer *et al* 1999; Stoneking 2000). However, it is possible that some errant sites are causing this effect and their removal from the dataset might result in a vast improvement in branching support. To identify these sites, we assumed the tree reconstructed with coding region sequences was the “true tree” and estimated how many individual substitutions would need to occur at each site in the D-loop dataset for the data to fit this tree. In this way, we worked backward and inferred the sites and sequences where parallel and back mutations must have occurred. We found that of the 80 parsimony informative sites (variants that are present in more than one sequence) in the D-loop dataset, 57 of these were present independently in multiple lineages. This means that the same substitution has occurred at the same site more than once in the dataset at over 70% of the informative sites. Therefore, the majority of the polymorphism found in the mitochondrial D-loop is not useful in inferring genetic relationships between sequences.

With the removal of the 44 sites from the D-loop dataset that were most strongly affected by homoplasy, the D-loop is forced to behave in a clock-like manner (Appendix 2, Figure 5). However, this does not result in a tree with better branching support as there are no longer sufficient polymorphic sites in that region with which to build an informative tree. When these 44 sites are removed from the dataset of complete mitochondrial genomes, some improvement is seen relative to the branching support of trees reconstructed with the coding region sequences or with the complete mitochondrial sequences (Appendix 2, Figure 6).

Although RFLP data provide genotypic information for sites in both the D-loop and the mitochondrial coding region, these data are also based on fragmentary information. In an attempt to assess the influence of building mtDNA phylogenies based on RFLP data alone, we inferred the nucleotide position at which the commonly used restriction enzymes would cleave the mtDNA strands. This information allowed us to catalogue
all polymorphism that could theoretically be identified using restriction enzymes. We found that in general, RFLP sites that have been identified as lineage specific only represent a portion of the informative sites that can be identified with complete sequence data and will therefore provide less support for tree topology. While RFLP data can define some major clades, the complete genome data provide more topological support and also distinguish a number of additional clades. Also, the availability of a dataset of complete mitochondrial genomes has allowed a more detailed analysis of the pattern of selection across the mitochondrial coding region. All protein coding genes showed an excess of synonymous substitutions compared to nonsynonymous substitutions consistent with the action of purifying selection (Appendix 2, Figure 2). Analysis of the substitution rate in the rRNA and tRNA genes reveals that the action of purifying selection is also preserving their structure and function.
Due to lower sea levels, mainland Australia, Tasmania and New Guinea have been joined several times over the last 70 thousand years, creating a landmass called Sahul. At the same times, Borneo, Sumatra and Java have been connected to the mainland of South East Asia forming the Sunda peninsula. These periods of low sea level would facilitate human migration from Southeast Asia to Sahul. If Australia and New Guinea were joined at the time of colonisation, it is possible that there would be genetic similarities between their native populations. Previous studies of mitochondrial D-loop variation (Redd and Stoneking 1999; van Holst Pellekaan et al 1998) and RFLP data (Stoneking et al 1990) have returned mixed results on the question of common origin. An analysis of Y chromosome data from Australia and Melanesia has suggested separate origins of Australian and New Guinean aborigines (Kayser et al 2001), which is in conflict with a study of autosomal loci (Roberts-Thomson et al 1996). In Papers I and II, we showed that superior data quantity and quality can be obtained from complete mitochondrial sequences, relative to the mitochondrial D-loop, and this could provide the basis for a stronger conclusion on the origin of Aboriginal Australians.

We obtained complete mitochondrial genome sequence data for 20 Australian Aborigines and 21 New Guineans, as well as 60 complete sequences from the peoples of Africa, India, Europe, Asia, Melanesia and Polynesia†. A Neighbor-Joining tree (Saitou and Nei 1987) reconstructed with the 101 coding region sequences shows a well-supported topology (Appendix 3, Figure 2). Non-African sequences are represented on two branches, indicated in the figure by dashed lines. All but one of the Australian sequences are present on branch 1, together with more than half the New Guinean sequences, all the European and Polynesian sequences and several Asian sequences. Our data indicate that some Australian sequences do show a closer relationship to some New Guinean sequences, particularly to those from the New Guinea highlands. This result is in line with the theory that the highlands of New Guinea became a refuge for the original inhabitants of that region when new colonists arrived and inhabited the coastal areas (Bellwood 1990). Because genetic divergence is expected to precede population divergence, modern humans must have arrived in Sahul since the time at which all the lineages on branch 1 diverged (71± 12 kya). A pairwise mismatch distribution for the Australian sequences shows a pattern indicative of a period of population growth (Appendix 3, Figure 3a), confirmed by tests of selective neutrality (Fu and Li 1989; Tajima 1989). The time that has passed since this expansion is dated from an estimate of Tau to be about 40 kya. Significantly negative Tajima’s $D$ values and unimodal mismatch distributions can result after a range expansion followed by a high rate of emigration to neighbouring demes (Ray et al 2003). It is therefore possible that this evidence of population expansion may instead indicate a spatial expansion as a relatively small number of people moved into a vast and unpopulated continent. Since it is unlikely that Australia was colonised by a very large heterogenous population, this expansion would have presumably taken place.

† of these 101 sequences, 49 were generated as part of a previous study (Ingman et al 2000) and 52 were sequenced for this study.
after the migration. In this way, we can frame the timing of human arrival in Australia between the coalescence date and the expansion date: 40 to 70 thousand years ago. This is in agreement with archaeological evidence which dates two human occupation sites in northern Australia (Roberts et al. 1990; Roberts et al. 1994) and one in the southeast region of Australia (Thorne et al. 1999; Bowler et al. 2003) at 50 to 60 kya, and evidence of animal extinctions across the continent, presumably as a result of hunting, about 46 kya (Roberts et al. 2001).

Consistent with previous investigations of mitochondrial sequence data (van Holst Pellekaan et al. 1998; Redd and Stoneking 1999), our study revealed high genetic diversity within both Australian and New Guinean populations. This is, however, disparate to the level of genetic diversity reported for the Y chromosome. Previous studies have found a unique Australia-specific Y chromosome haplotype in about 50% of Australian Aboriginal males (Kayser et al. 2001; Redd et al. 2002). Kayser and colleagues (2001) suggested that the high frequency of this haplotype in Australia could be the result of a relatively recent population expansion. However, the bottleneck among other lineages that must have accompanied such a scenario is not evident among mitochondrial sequences. The relatively high genetic diversity among mitochondrial genomes could be attributed to a combination of high substitution rate (Brown et al. 1979), patrilocality (Seielstad et al. 1998; Oota et al. 2001) and polygyny (Birdsell 1993).
Paper IV – Ancient human migrations

While the broad outline of modern human origins has been established through analysis of physical and molecular data, important details, such as migrations and the relationships of global populations, remain to be fixed. Many population genetics studies suffer from the complexities of human populations, partly due to admixture. Large-scale population based analyses allow greater resolution than those based on small numbers of sequences and may yield valuable information on the history of human populations. In this manuscript we have analysed large SNP and sequence data sets in order to increase our understanding of the fundamental associations of extant peoples.

The general phylogenetic structure of a global selection of complete human mitochondrial genomes has been established (Ingman et al 2000). The deepest branches are reserved for sub-Saharan African lineages and all non-African sequences are present on two main branches that diverged from African lineages in the last 100 thousand years (Appendix 1, Figure 2). The sequences on these two non-African branches are delineated by 7 nucleotide substitutions. In order to study the evolution and spread of non-Africans, we typed these SNPs in a global selection of DNA from 1759 individuals. The human genome diversity panel (HGDP) consists of 1064 DNA samples from 52 world populations (Cann et al 2002) and we supplemented this number with extra samples obtained separately to better represent African and Asian diversity.

These 7 SNPs identify 14 haplotypes among these samples (Appendix 4, Figure 1). While 7 of these 14 haplotypes are found in Africa, only two major haplotypes are found in all African populations at high frequencies. These two haplotypes are absent in all the non-African populations examined except the Middle East (Israel) and South Asia (Pakistan) (Appendix 4, Table 1). In Europe, East Asia, Australia and the Americas, two different haplotypes predominate with very few other haplotypes present at appreciable frequencies. The presence of only two major mtDNA haplotypes in most non-African populations can be attributed to a population bottleneck associated with the exodus of modern humans from Africa and their spread throughout the world. The presence of African haplotypes in the Middle East and Pakistan is evidence of a genetic connection between these populations possibly through repeated human movements.

In order to add information to the scenario outlined by the SNP data, we have sequenced more than 300 mitochondrial coding region sequences of 6300 base pairs in length. These sequences derive from samples in the CEPH human genome diversity panel (China, Pakistan, Israel, Mexico, Japan and Yakut) as well as our own sample collection (Thai hill tribes, Japan and Yakut). A population tree reconstructed with these and other published sequences shows a basal split between East Asian and South Asian populations. Using various standard tests (Tajima 1989, Fu and Li 1993), significant departures from neutrality were detected in several of the populations. This can be interpreted as changes in the pattern of polymorphism in a population due to a demographic or spatial expansion. The amount of time that has passed since each expansion occurred was calculated. The oldest expansion (52.5 kya) is among the
lineages of present day Chinese people with other closely related populations having more recent expansions (Appendix 4, Table 2). This could indicate that the China region was colonised first and a subsequent demographic or spatial expansion seeded migrations to other regions. The expansion times on the other non-African branch are more recent. The population tree and expansion analyses indicate that peoples of Israel and Europe are younger than their South Asian contemporaries. The data are consistent with a recolonisation of the Middle East by a back migration from South Asia and a subsequent migration to Europe.

The evidence of population expansions, whether they are demographic or spatial, could be associated with the founding of new populations in uninhabited territories. However, while the cause of these expansions can never definitely be known, it is also possible that some sort of cultural adaptation occurred that allowed human settlements to support higher population densities or facilitated dispersal to new territories.
Concluding remarks

Analysis of mitochondrial DNA has been a potent tool in the understanding of human evolution and migration. The data, analyses and conclusions we have presented in these publications reveal the importance of complete mitochondrial DNA sequences in this domain. Paper I shows that the use of mitochondrial genomes can provide increased phylogenetic resolution and cleaner analyses than had previously been published using partial mtDNA sequences. Although other genetic data have already indicated that a recent African origin of modern humans was likely, the recognition of this within the bounds of statistical significance adds weight to the hypothesis that all living humans share a common ancestor within the last 200 thousand years.

Paper II discusses the extensive variation in substitution rate found in the mitochondrial D-loop and what effect the resultant homoplasy may have on phylogenetic inference. The results show that complete mitochondrial genome analysis will add important information to studies of human evolution when used exclusively, or in conjunction with other datasets. While the analyses presented in this manuscript convey damning evidence against the inherent problems with D-loop analysis, one should remember that even though those studies have not achieved many statistically sound results, they have still provided important information. It is expected that many results based on D-loop data will be verified in the future by other means.

Paper III provides evidence of a genetic link between Australian Aborigines and the indigenous people of the New Guinea highlands. The date put forward for the initial colonisation of Australia is between 40 and 70 thousand years ago, in concert with archaeological evidence. There may be substantial genetic variation between regions/tribal groups and any future comprehensive study should involve hundreds if not thousands of samples from a wide range of locales and typing from multiple genetic systems.

Paper IV examines an array of SNPs in a global set of 1759 DNA samples. The pattern of diversity underlines the genetic bottleneck associated with the African exodus and modern human migration throughout the rest of the world. An ancient connection between African populations and those nearby is also indicated. A study of more than 800 mitochondrial sequences points to a divergence among non-African peoples beginning more than 50 thousand years ago, possibly associated with differential migrations. The data indicate that the contemporary peoples of the Middle East and Europe could result from a back migration from South Asia.

The course of evolution to modern humans is complex. The fossil record is incomplete and, unfortunately, technical problems with time travel currently exist. The best we can do is to integrate the various evidence as well as possible and try to keep an open mind. Becoming human was a gradual process of accumulation of genotypic and phenotypic changes in all genetic systems as well as behavioural and cultural advances over many thousands of years. Because the different genetic systems can have different effective population sizes, the time to the common genetic ancestor of all extant human lineages varies with the locus in question. If we can assume that the phenotypic change that tipped the balance occurred sometime between the estimates
for the most recent common ancestor of the different loci, then we can only guess that this happened sometime in the last million years. The mitochondrial genome is only one locus and can only reflect the history of females in a population. In the future, many studies will be presenting both mitochondrial and nuclear data together for a balanced view of human evolution and population histories. The Human Genome Project, coupled with studies of sequence data from regions with low recombination on autosomes and on the X chromosome, and in-depth analyses of genetic variation on the paternally inherited Y chromosome, are uncovering a large amount of polymorphism that is of vital importance in the study of human evolution. With new technology regularly surpassing previous methods, the small size of the mitochondrial genome that so attracted researchers in the beginning will become a significant limiting factor. It is now possible to examine many thousands of sites for variants in a single assay rendering the data available from mitochondria insignificant in comparison to the vast nuclear genome. In view of the fact that more SNPs are constantly becoming available, the greater geographic variability coupled to the stability of binary polymorphisms in the Y chromosome will see this locus rise to the forefront, offering a finer resolution of human population history. As human population genetics becomes an important cornerstone in the field of medical genetics, the immediate limiting factor will be methods of analysis rather than data production.
ACKNOWLEDGEMENTS

First I would like to thank Ulf, my supervisor, who had enough faith to hand me the reins and the acumen to know when to take them back.

The girls of the Uppsala Genome Centre (Inger, Jenny and Ann-Sofi) provided vital help and, for the most part, allowed me to use their equipment without restriction. Thanks for putting up with me and helping me though this.

I also want to thank Per-Ivan who was never too busy to provide invaluable computer help that I’m sure came at a cost to his own work.

Thank you to everyone at Medical Genetics for helping to make these years happy and productive.

I have not forgotten the early tolerance and encouragement from James and Mandy that allowed me to follow this path.

This work was supported by grants from the Swedish Natural Sciences Research Council and the Beijer Foundation.
REFERENCES


Wright, S. Evolution in Mendelian populations. *Genetics* 16, 97-159 (1931).

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to October, 1985, the series was published under the title “Abstracts of Uppsala Dissertations from the Faculty of Medicine”.)