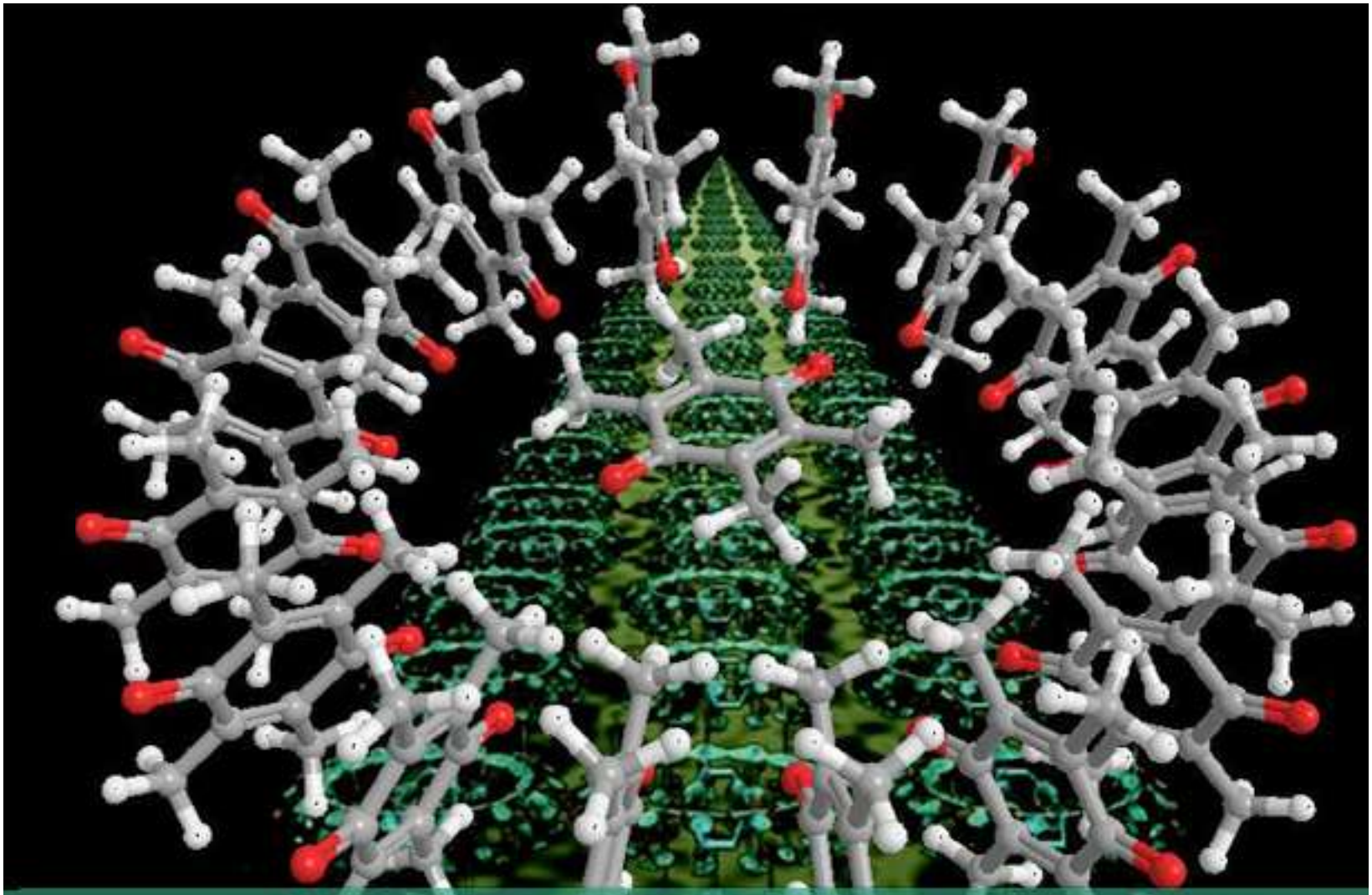


Molecular Anthropology

Joye Frick

First Edition, 2012

ISBN 978-81-323-3155-1

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Molecular Anthropology

Molecular anthropology is a field of anthropology in which molecular analysis is used to determine evolutionary links between ancient and modern human populations, as well as between contemporary species. Generally, comparisons are made between sequence, either DNA or protein sequence, however early studies used comparative serology.

By examining DNA sequences in different populations, scientist can determine the closeness relationships between populations (or within populations). Certain similarities in genetic makeup let molecular anthropologists determine whether or not different groups of people belong to the same haplogroup, and thus if they share a common geographical origin. This is significant because it allows anthropologists to trace patterns of migration and settlement, which gives helpful insight as to how contemporary populations have formed and progressed over time.

Molecular anthropology has been extremely useful in establishing the evolutionary tree of humans and other primates, including closely related species like chimps and gorillas. While there are clearly many morphological similarities between humans and chimpanzees, for example, certain studies also have concluded that there is roughly a 98 percent commonality between the DNA of both species. However, more recent studies have modified the commonality of 98 percent to a commonality of only 94 percent, showing that the genetic gap between humans and chimps is bigger than originally thought. Such information is useful in searching for common ancestors and coming to a better understanding of how humans evolved.

Haploid Loci in molecular anthropology

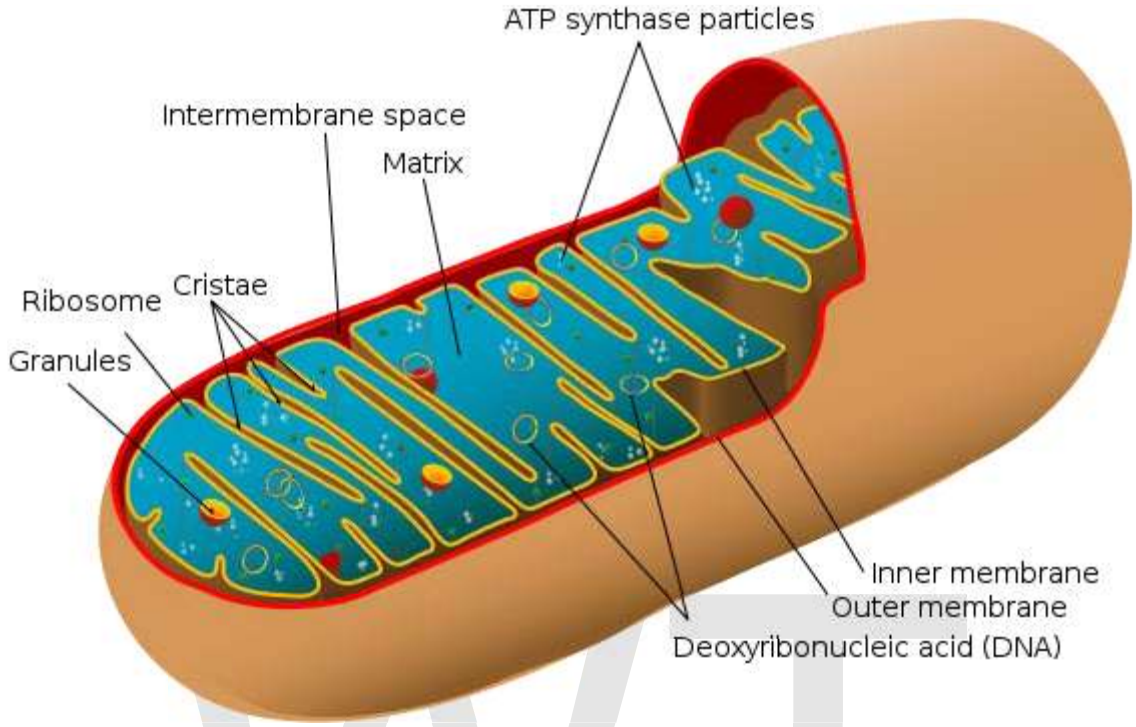


Image of mitochondrion. There many mitochondria within a cell, and DNA in them replicate independently of the chromosomes in the nucleus.

There are two continuous linkage groups in human that are carried by a single sex. The first is the Y-chromosome, which is passed from father to son. Rarely do anatomical females carry a Y chromosome as a result of genetic defect. The other linkage group is the mitochondrial DNA (mtDNA). MtDNA can only be passed to the next generation by females but only under highly exceptional circumstances is mtDNA passed through males. The non-recombinant portion of the Y chromosome and the mtDNA, under normal circumstances, do not undergo productive recombination. Part of Y chromosome can undergo recombination with X chromosome and within ape history the boundary has changed. Such recombinant changes in the non-recombinant region of Y are extremely rare.

Mitochondrial DNA

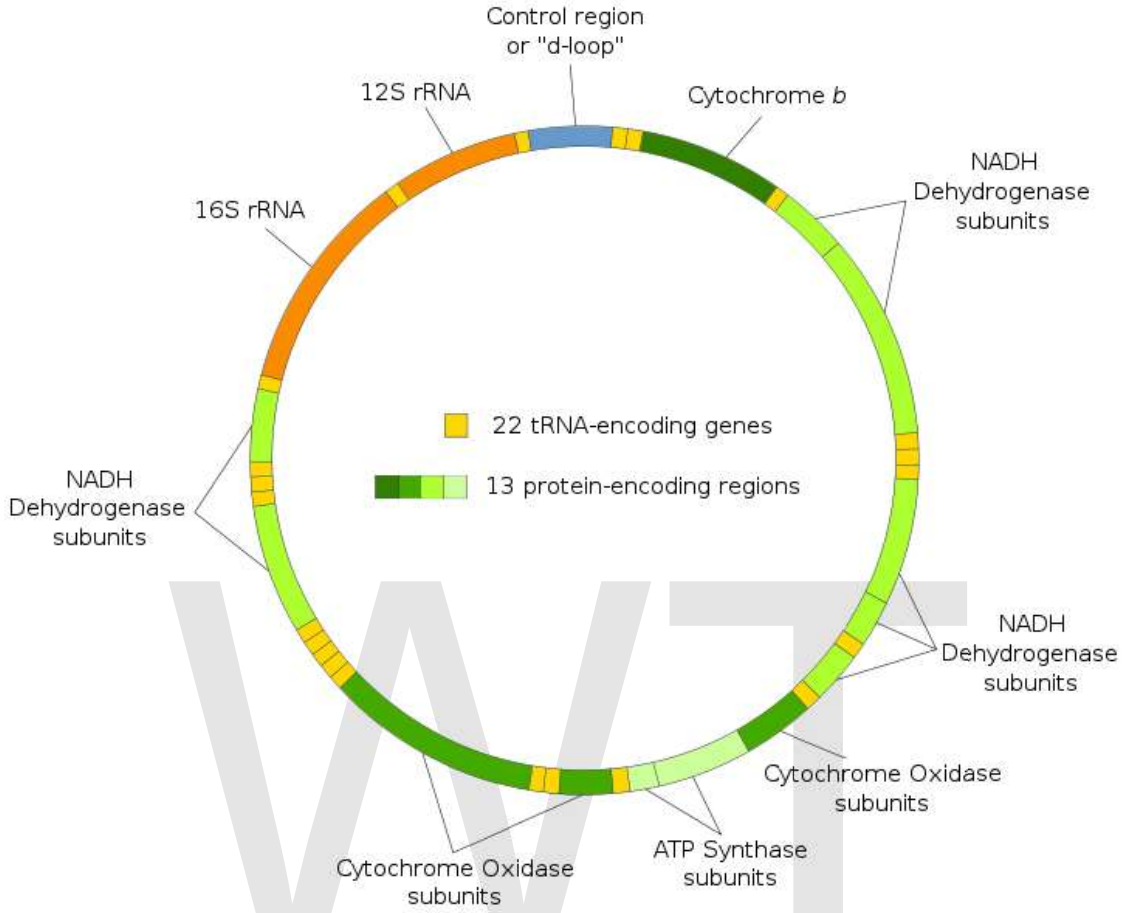
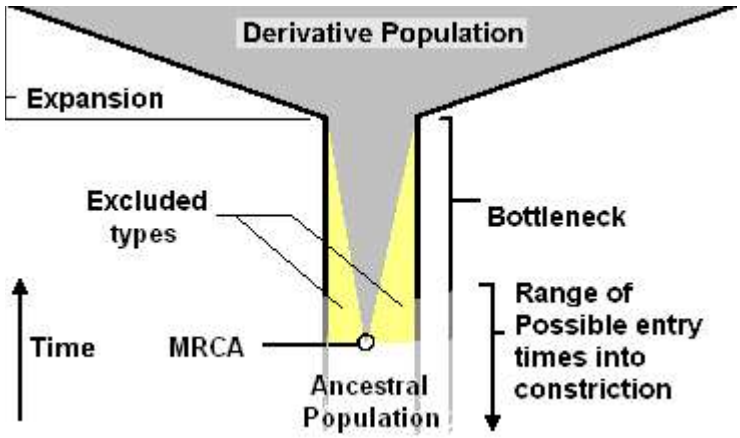


Illustration showing mitochondrial DNA, the control region (D-loop, hypervariable regions 1 and II are located on left and right side at the top, respectively)

Mitochondrial DNA became an area of research in phylogenetics in the late 1970s. Unlike genomic DNA is offered advantages in that it did not undergo recombination. The process of recombination, if frequent enough corrupts the ability to create parsimonious trees because stretches of amino acid subsititions (SNPs). When looking between distantly related species, recombination is less of a problem since recombination between branches from common ancestors is prevented after true speciation occurs. When examining closely related species, or branching within species recombination creates a large number of 'irrelevant SNPs' for cladistic analysis. MtDNA, through the process of organelle division, become clonal over time, very little, too often none, of that paternal mtDNA is passed. While recombination may occur in mtDNA, there is little risk that it will be passed to the next generation. As a result mtDNA become clonal copies of each other, except when a new mutation arises. As a result mtDNA does not have pitfalls of autosomal loci when studied in interbreeding groups. Another advantage of mtDNA is that the hyper-variable regions evolve very quickly, this exhibits that certain regions of mitochondrial DNA approach neutrality. This allowed the use of mitochondrial DNA to determine that the relative age of the human population was small having gone through a

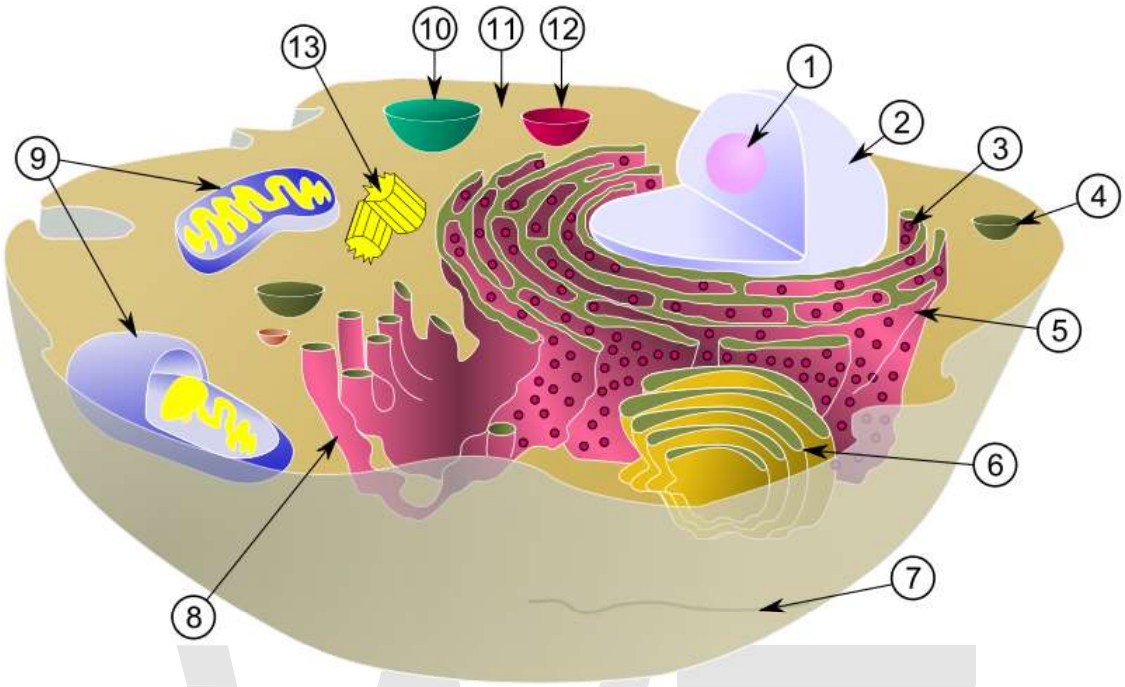
recent constriction at about 150,000 years ago. Mitochondrial DNA has also been used to verify the proximity of chimpanzees to humans relative to gorilla, and to verify the relationship of these 3 species relative to orangutan.



A population bottleneck, as illustrated was detected by intrahuman mtDNA phylogenetic studies, the length of the bottleneck itself is indeterminant per mtDNA.

More recently the mtDNA genome has been used to estimate branching patterns in peoples around the world, such as when the new world was settled and how. The problem with these studies have been that they rely heavily on mutations in the coding region. Researchers have increasingly discovered that as humans moved from Africas south-eastern regions, that more mutations accumulated in the coding region than expected, and in passage to the new world some groups are believe to have passed from the Asian tropics to Siberia to an ancient land region called Beringia and quickly migrating to south America. Many of the mtDNA have far more mutations and at rarely mutated coding sites relative to expectations of neutral mutations.

Mitochondrial DNA offers another advantage over autosomal DNA. There are generally 2 to 4 copies of each chromosome in each cell (1 to 2 from each parent chromosome). For mtDNA there can be dozens to hundreds in each cell. This increases the amount of each mtDNA loci by at least a magnitude. For ancient DNA, in which the DNA is highly degraded, the number of copies of DNA is helpful in extending and bridging short fragments together, and decreases the amount of bone extracted from highly valuable fossil/ancient remains. Unlike Y chromosome both male and female remains carry mtDNA in roughly equal quantities.



Schematic of typical animal cell, showing subcellular components. Organelles: (1) nucleolus (2) nucleus (9) mitochondria

Y chromosome

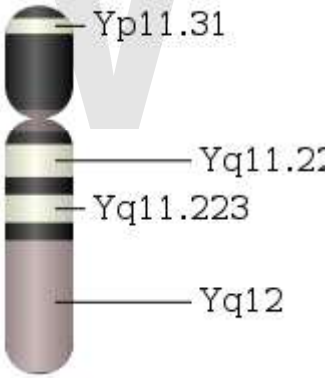


Illustration of human Y chromosome

Y chromosome is found in the nucleus of normal cells (Nuclear DNA). Unlike mtDNA, it has mutations in the non-recombinant portion (NRY) of the chromosome spaced widely apart, so far apart that finding the mutations on new Y chromosomes is labor intensive relative to mtDNA. Many studies rely on tandem repeats; however, tandem repeat can expand and retract rapidly and in some predictable patterns. Y chromosome only tracks male lines, and is not found in females; whereas mtDNA can be traced in males even though they fail to pass mtDNA. In addition it has been estimated that effective male populations in the prehistoric period were typically 2 females per male, and recent studies show that cultural hegemony plays a large role in the passage of Y. This has created

disconcordance between the time to most recent common ancestor of males and females. The estimates for Y TMRCA range from 1/4th to less than 1/2 that of mtDNA TMRCA. It is unclear whether this is due to high male-to-female ratios in the past coupled with repeat migrations from Africa, as a result of mutational rate change, or some have even proposed that females of the LCA between chimps and humans continued to pass DNA millions after males ceased to pass DNA. At present the best evidence suggests that in migration the male to female ratio in humans may have declined causing a trimming of Y diversity on multiple occasions within and outside of Africa.

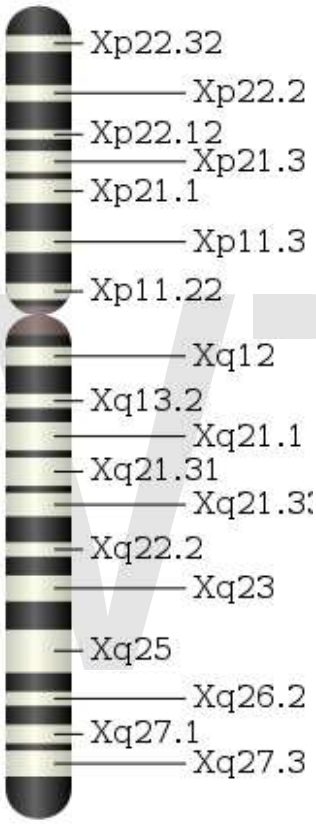


Diagram of human X chromosome showing genetic map

For short range molecular phylogenetics and molecular clocking Y chromosome is highly affective and creates a second perspective. One argument that arose was that the Maori by mtDNA appear to have migrated from Eastern China or Taiwan, by Y chromosome from Papua New Guinea region. When HLA haplotypes were used to evaluate the two hypothesis it was uncovered that both were right, that the Maori were an admixed population. Such admixtures appear to be common in the human population and thus the use of a single haploid loci can give a biased perspective.

X-linked Studies

The X-chromosome is also a form of nuclear DNA. Since it is found as 1 copy in males and 2 non-identical chromosomes in females it has a ploidy of 1.5. However, in humans the effective ploidy is somewhat higher, ~1.7, as females in the breeding population have tended to outnumber males by 2:1 during a large portion of human prehistory. Like mtDNA, X-linked DNA tends to over emphasize female population history much more than male. There have been several studies of loci on X chromosome, in total 20 sites have been examined. These include PDHA1, PDHA1, Xq21.3, Xq13.3, Zfx, Fix, Il2rg, Plp, Gk, Ids, Alas2, Rrm2p4, AmeIX, Tnfsf5, Licam, and Msn. The time to most recent common ancestor(TMRCA) ranges from fixed to ~1.8 million years, with a median around 700ky. These studies roughly plot to the expected fixation distribution of alleles, given linkage disequilibrium between adjacent sites. For some alleles the point of origin is elusive, for others, the point of origin points toward Sub-saharan Africa. There are some distinctions within SSA that suggest a smaller region, but there is not adequate enough sample size and coverage to define a place of most recent common ancestor. The TMRCA is consistent with and extends the bottleneck implied by mtDNA, confidently to about 500,000 years.

Autosomal Loci

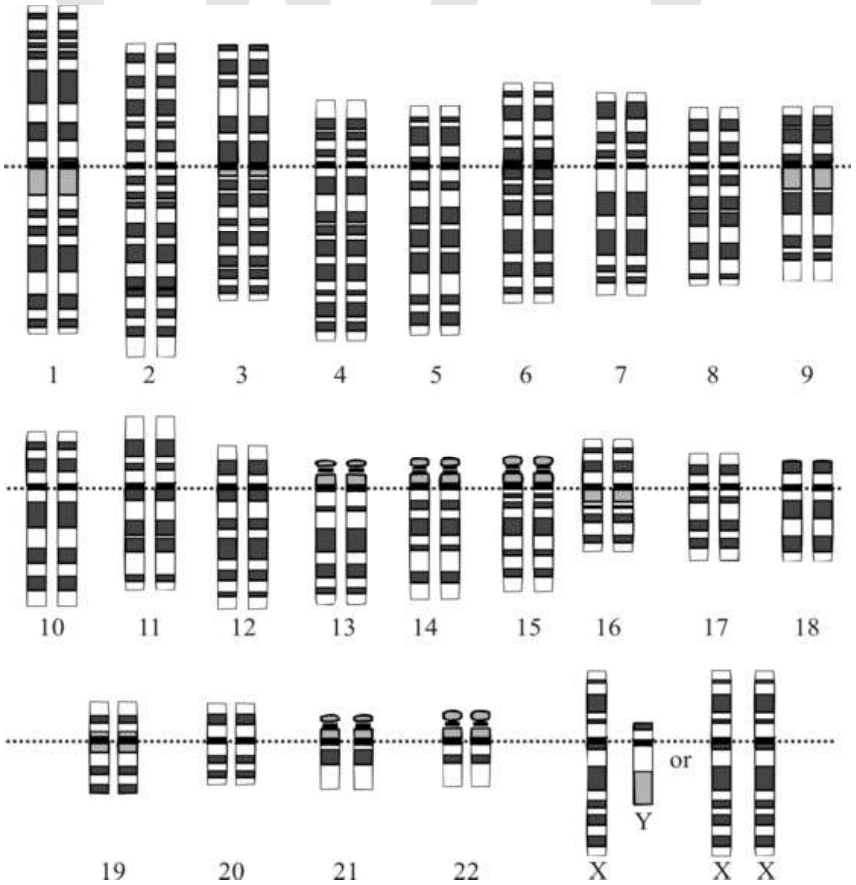
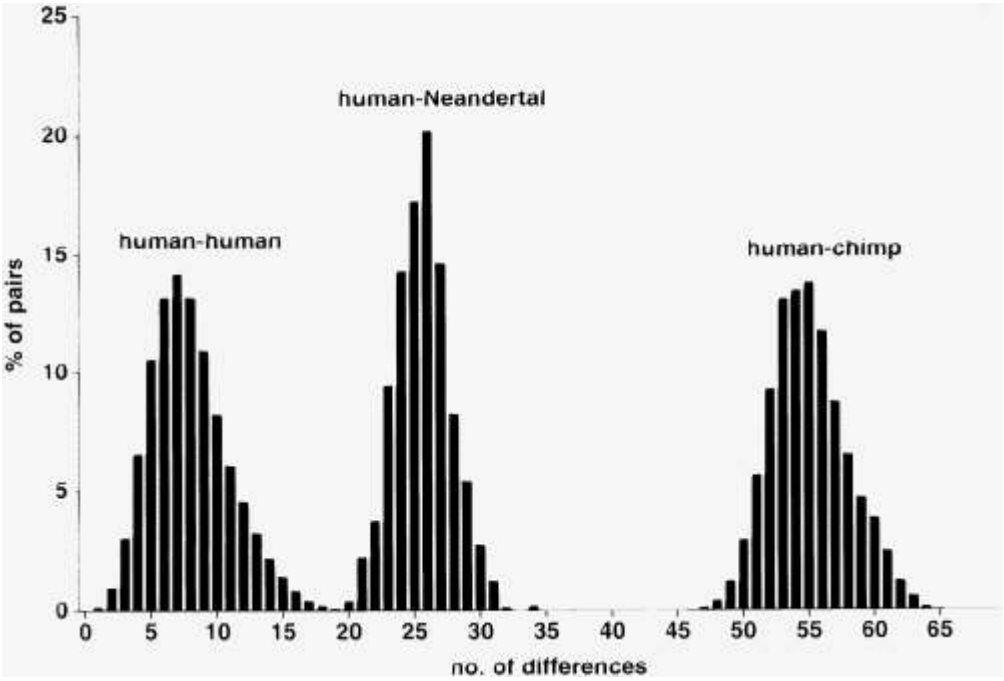


Diagram of human karyotype

Ancient DNA sequencing

Since Krings Neandertal mtDNA have been sequenced, and the sequence similarity indicates an equally recent origin from a small population on the Neanderthal branch of late hominids. MCR1 gene has also been sequenced but the results are controversial, with one study claiming that contamination issues cannot be resolved from human Neandertal similarities. Critically however no DNA sequence has been obtained from Homo erectus, Homo floriensis, or any of the other late hominids. Some of the ancient sequences obtained have highly probable errors, and proper control to avoid contamination.



Comparison of differences between human and Neanderthal mtDNA

Causes of errors

The molecular phylogenetics is based on quantification substitutions and then comparing sequence with other species, there are several points in the process which create errors. The first and greatest challenge is finding "anchors" that allow the research to calibrate the system. In this example, there are 10 mutations between chimp and humans, but the researcher has no known fossils that are agreeably ancestral to both but not ancestral to the next species in the tree, gorilla. However, there are fossils believed to be ancestral to Orangutans and Humans, from about 14 million years ago. So that the researcher can use Orangutan and Human comparison and comes up with a difference of 24. Using this he can estimate (24/(14*2, the "2" is for the length of the branch to Human (14my) and the branch to Orangutan (14 my) from their last common ancestor (LCA). The mutation rate at 0.857 for a stretch of sequence. Mutation rates are given, however, as rate per nucleotide(nt)-site, so if the sequence were say 100 nt in length that rate would be 0.00857/nt per million years. Ten mutations*100nt/(0.00857*2) = 5.8 million years.

The problem of calibration

There are several problems not seen in the above. First, mutations occur as random events. Second, the chance that any site in the genome varies is different from the next site, a very good example is the codons for amino acids, the first two nt in a codon may mutate at 1 per billion years, but the third nt may mutate 1 per million years. Unless scientist study the sequence of a great many animals, particularly those close to the branch being examined, they generally do not know what the rate of mutation for a given site. Mutations do occur at 1st and 2nd positions of codons, but in most cases these mutations are under negative selection and so are removed from the population over small periods of time. In defining the rate of evolution in the anchor one has the problem that random mutation creates. For example a rate of .005 or .010 can also explain 24 mutations according to the binomial probability distribution. Some of the mutations that did occur between the two have reverted, hiding an initially higher rate. Selection may play into this, a rare mutation may be selective at point X in time, but later climate may change or the species migrates and it is not longer selective, and pressure exerted on new mutations that revert the change, and sometimes the reversion of a nt can occur, the greater the distance between two species the more likely this is going to occur. In addition, from that ancestral species both species may randomly mutate a site to the same nucleotide. Many times this can be resolved by obtaining DNA samples from species in the branches, creating a parsimonious tree in which the order of mutation can be deduced, creating branch-length diagram. This diagram will then produce a more accurate estimate of mutations between two species. Statistically one can assign variance based on the problem of randomnicity, back mutations, and parallel mutations (homoplasies) in creating an error range.

There is another problem in calibration however that has defied statistical analysis. There is a true/false designation of a fossil to a least common ancestor. In reality the odds of having the least common ancestor of two extant species as an anchor is low, often that fossil already lies in one branch (underestimating the age), lies in a third branch (underestimating the age) or in the case of being within the LCA species, may have been millions of years older than the branch. To date the only way to assess this variance is to apply molecular phylogenetics on species claimed to be branch points. This only, however identifies the 'outlying' anchor points. And since it is more likely the more abundant fossils are younger than the branch point the outlying fossil may simply be a rare older representative. These unknowns create uncertainty that is difficult to quantify, and often not attempted.

Recent papers have been able to estimate, roughly, variance. The general trend as new fossils are discovered, is that the older fossils underestimated the age of the branch point. In addition to this dating of fossils has had a history of errors and there have been many revised datings. The age assigned by researchers to some major branch points have almost doubled in age over the last 30 years. An excellent example of this is the debate over LM3 (Mungo lake 3) in Australia. Originally it was dated to around 30 ky by carbon dating, carbon dating has problems, however, for sampled over 20ky in age, and severe

problems for samples around 30ky in age. Another study looked at the fossil and estimated the age to be 62 ky in age.

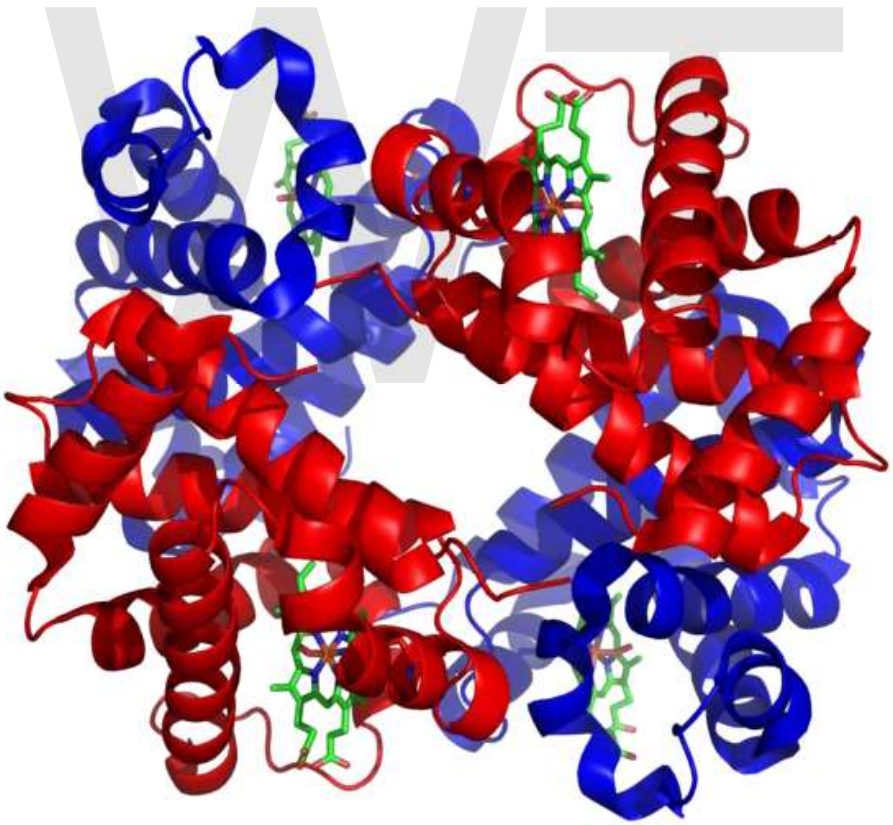
At the point one has an estimation of mutation rate, given the above there must be two sources of variance that need to be cross-multiplied to generate an overall variance. This is infrequently done in the literature.

Problems in estimating TMRCA

Time to most recent common ancestor (**TMRCA**) combines the errors in calibration with errors in determining the age of a local branch.

History of Molecular Anthropology

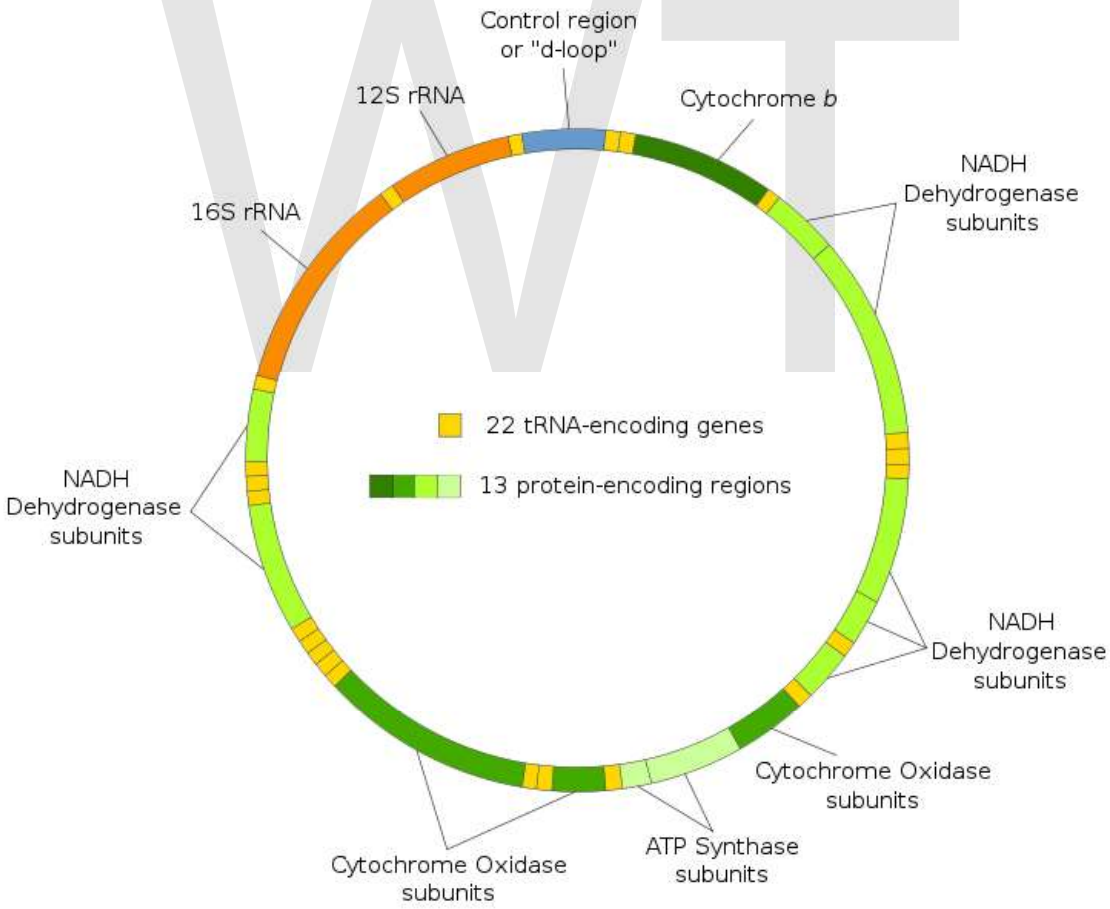
The protein era



Structure of human hemoglobin. Hemoglobins from dozens of animals and even plants were sequenced in the 1960s and early 1970s

With DNA newly discovered as the genetic material, in the early 1960s protein sequencing was beginning to take off. Protein sequencing began on cytochrome C and Hemoglobin. Gerhard Braunitzer sequenced hemoglobin and myoglobin, in total more than hundreds of sequences from wide ranging species were done. In 1967 A.C. Wilson began to promote the idea of a "molecular clock". By 1969 molecular clocking was applied to anthropoid evolution and V. Sarich and A.C. Wilson found that albumin and hemoglobin has comparable rates of evolution, indicating chimps and humans split about 4 to 5 million years ago. In 1970, Louis Leakey confronted this conclusion with arguing for improper calibration of molecular clocks. By 1975 protein sequencing and comparative serology combined were used to propose that humans closest living relative (as a species) was the chimpanzee. In hindsight, the last common ancestor (LCA) from humans and chimps appears to older than the *Sarich and Wilson* estimate, but not as old as Leakey claimed , either. However, Leakey was correct in the divergence of old and new world monkeys, the value Sarich and wilson used was a significant underestimate. This error in prediction capability highlights a common theme.

The DNA era

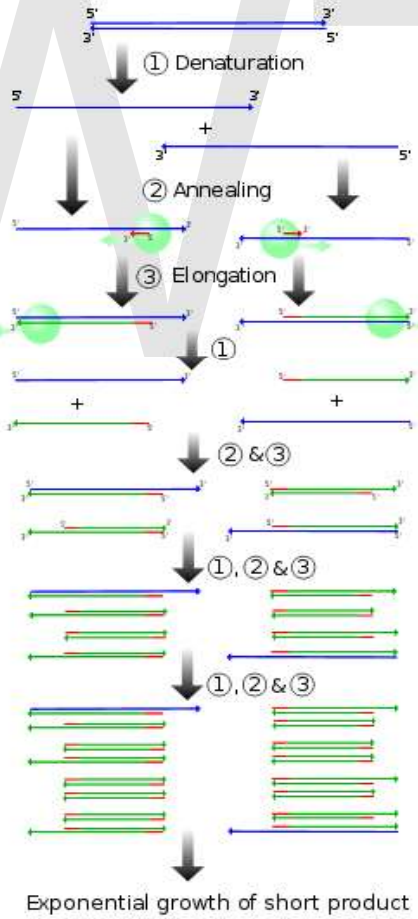


Restriction fragment length polymorphisms studies the cutting of mtDNA into fragements, Later the focus of PCR would be on the D 'contol'-loop, at the top of the circle

RLFP and DNA hybridization

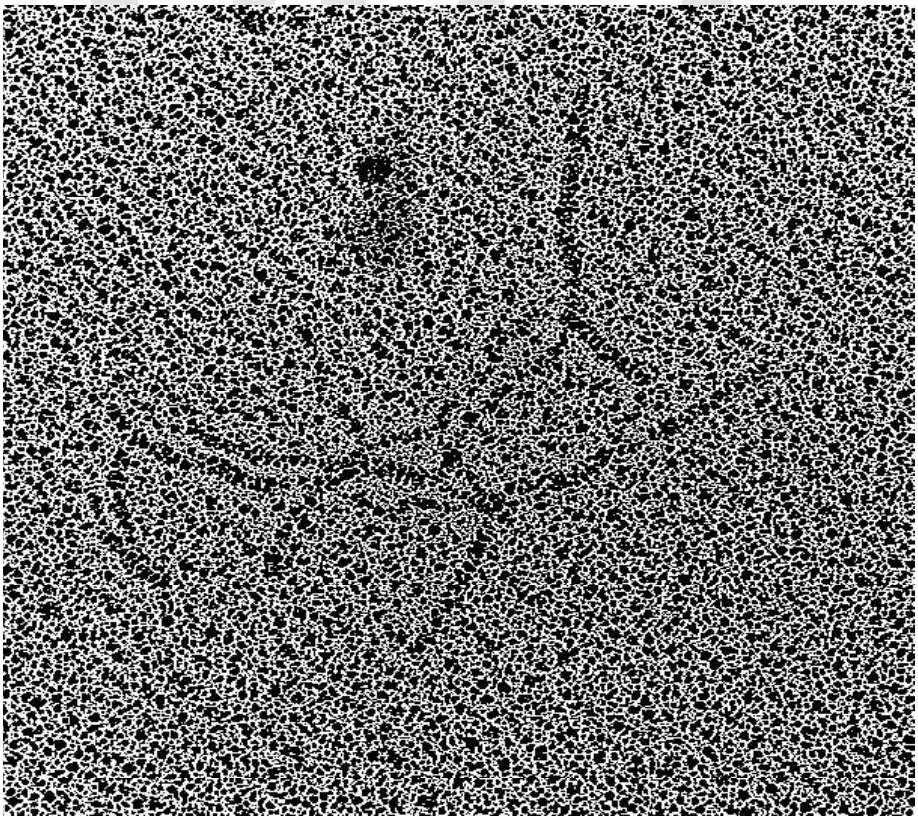
In 1979 W.M.Brown and Wilson began looking at the evolution of mitochodrial DNA in animals, and found they were evolving rapidly. The technique they used was restriction fragment length polymorphism(**RFLP**) which was more affordable at the time compared to sequencing. In 1980, W.M. Brown, looking at the relative variation between human and other species, recognizes there was a recent constriction (180,000 years ago) in the human population. A year later Brown and Wilson were looking at RFLP fragments and determined the human population expanded more recently than other ape populations. In 1984 the first DNA sequence from an extinct animal was done. Sibley and Ahlquist apply DNA-DNA hybridization technology to anthropoid phylogeny, and see pan/human split closer than gorilla/pan or gorilla/human split, a highly controversial claim. However, in 1987 they were able to support their claim. In 1987, Cann, Stoneking and Wilson suggest, by RFLP analysis of human mitochondrial DNA, that humans evolved from a constrict in Africa of a single female in a small population, ~10,00 individuals, 200,000 years ago.

The era of PCR



PCR could rapidly amplify DNA from 1 molecule to billions allowing sequencing from human hairs or ancient DNA

In 1987, PCR-amplification of mtDNA was first used to determine sequences. In 1991 Vigilante et al. published the seminal work on mtDNA phylogeny implicating sub-saharan Africa as the place of humans most recent common ancestors for all mtDNAs. The war between out-of-Africa and multiregionalism, already simmering with the critiques of Allan Templeton, soon escalated with the paleoanthropologist, like Milford Wolpoff, getting involved. In 1995, F. Ayala published his critical Science article 'The Myth about Eve', which relied on HLA-DR sequence. At the time, however Ayala was not aware of rapid evolution of HLA loci via recombinatory process. In 1996, Parham and Ohta published their finds on the rapid evolution of HLA by short-distance recombination ('gene conversion' or 'abortive recombination'), weakening Ayala's claim (Parham had actually written a review a year earlier, but this had gone unnoticed). A stream of papers would follow from both sides, many with highly flawed methods and sampling. One of the more interesting was Harris and Hey, 1998 which showed that the TMCRA (time to most recent common ancestor) for the PDHA1 gene was well in excess of 1 million years. Given a ploidy at this locus of 1.5 (3 fold higher than mtDNA) the TMRCA was more than double the expectation. While this falls into the 'fixation curve' of 1.5 ploidy (Averaging 2 female and 1 male) the suggested age of 1.8 my is close a significantly deviant p-value for the population size, possibly indicating that the human population shrank or split off of another population. Oddly, the next X-linked loci they examined, Factor IX, showed a TMRCA of less than 300,000 years.



Cross-linked DNA extracted from the 4,000 year-old liver of an Ancient Egyptian priest Called Nekht-Ankh.

Ancient DNA

Ancient DNA sequencing had been conducted on a limited scale up to the late 1990s when the folks at the Max Plank Institute would shock the anthropology world by sequencing DNA from an estimated 40,000 year old Neanderthal. The results of that experiment is that the differences between humans living in Europe, many of which were derived from haplogroup H (CRS), Neandertals branched from humans more than 300,000 years before haplogroup H reached Europe. While the mtDNA and other studies continued to support a unique recent African origin, this new study basically answered critiques from the Neandertal side.

Genomic Sequencing

Significant progress has been made in genomic sequencing since Ingman and colleague published their finding on mitochondrial genome. Several papers on genomic mtDNA have been published, there is considerable variability in the rate of evolution, and rate variation and selection are evident at many sites. In 2007 Gonder et al., proposed that a core population of humans, with greatest level of diversity and lowest selection once lived in the region of Tanzania and proximal parts of southern Africa, since humans left this part of Africa, mitochondria have been selectively evolving to new regions.

Critical Progress

Critical in the history of molecular anthropology:

- That molecular phylogenetics could compete with comparative anthropology for determining the proximity of species to humans.
- Wilson and King realized in 1975, that while there was equity between the level of molecular evolution branching from chimp to human to putative LCA, that there was an inequity in morphological evolution. Comparative morphology based on fossils could be biased by different rates of change.
- Realization that in DNA there are multiple independent comparisons. Two techniques, mtDNA and hybridization converge on a single answer, chimps as a species are most closely related to humans.
- The ability to resolve population sizes based on the 2N rule, proposed by Kimura in the 1950s. To use that information to compare relative sizes of population and come to a conclusion about abundance that contrasted observations based on the paleontological record. While human fossils in the early and middle stone age are far more abundant than Chimpazee or Gorilla, there are few unambiguous chimpanzee or gorilla fossils from the same period

Loci that have been used in molecular phylogenetics:

Cytochrome C
Serum Albumin
Hemoglobin - Braunitizer, 1960s, Harding et al. 1997

Mitochondrial D-loop - Wilson group, 1980, 1981, 1984, 1987, 1989, 1991(posthumously) - TMRCA about 170 kya.
Y-chromosome
HLA-DR - Ayala 1995 - TMRCA for locus is 60 million years.
CD4 (Intron) - Tishkoff, 1996 - most of the diversity is in Africa.
PDHA1 (X-linked) Harris and Hey - TMRCA for locus greater than 1.5 million years.

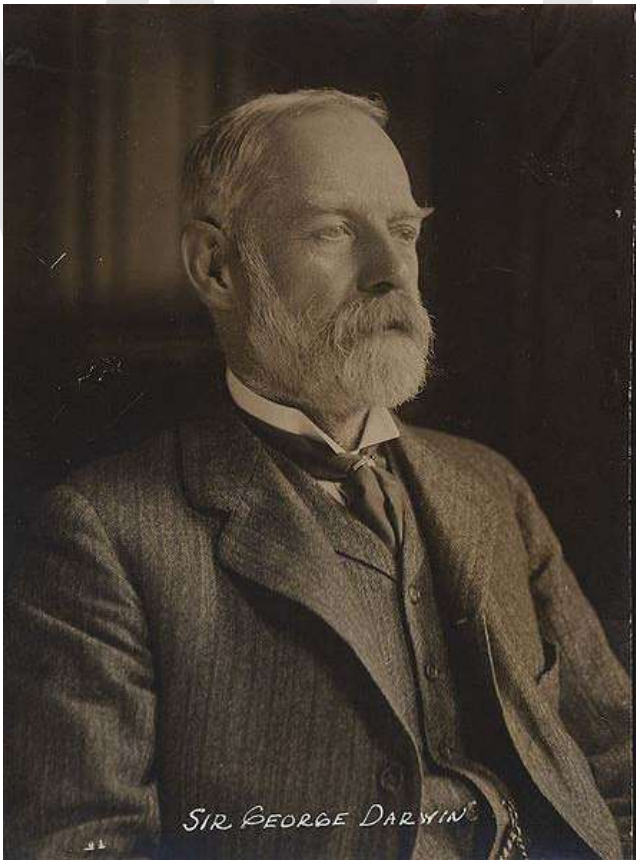
Xlinked loci: PDHA1, Xq21.3, Xq13.3, Zfx, Fix, Il2rg, Plp, Gk, Ids, Alas2, Rrm2p4, AmeIX, Tnfsf5, Licam, and Msn
Autosomal:Numerous.

Chapter- 2

Genetic Genealogy

Genetic genealogy is the application of genetics to traditional genealogy. Genetic genealogy involves the use of genealogical DNA testing to determine the level of genetic relationship between individuals.

History



George Darwin, son of Charles Darwin, was the first to estimate the frequency of first-cousin marriages

The investigation of surnames in genetics can be said to go back to George Darwin, a son of Charles Darwin. In 1875, George Darwin used surnames to estimate the frequency of first-cousin marriages and calculated the expected incidence of marriage between people of the same surname (isonymy). He arrived at a figure between 2.25% and 4.5% for cousin-marriage in the population of Great Britain, with the upper classes being on the high end and the general rural population on the low end. (His parents, Charles Darwin and Emma Wedgwood, were first cousins.) This simple study was innovative for its era. The next stimulus toward using genetics to study family history had to wait until the 1990s, when certain locations on the Y chromosome were identified as being useful for tracing male-to-male inheritance.

Dr. Karl Skorecki, a Canadian nephrologist of Ashkenazi parentage, noticed that a Sephardic fellow-congregant who was a Kohen like himself had completely different physical features. According to Jewish tradition, all Kohanim are descended from the priest Aaron, brother of Moses. Skorecki reasoned that if Kohanim were indeed the descendants of only one man, they should have a common set of genetic markers and should perhaps preserve some family resemblance to each other.

To test that hypothesis, he contacted Professor Michael Hammer of the University of Arizona, a researcher in molecular genetics and pioneer in Y chromosome research. Their report in the *Nature* in 1997 sent shock waves through the worlds of science and religion. A particular marker was indeed more likely to be present in Jewish men from the priestly tradition than in the general Jewish population. It was apparently true that a common descent had been strictly preserved for thousands of years. Moreover, the data showed that there were very few "non-paternity events".

The first to test the new methodology in general surname research was Bryan Sykes, a molecular biologist at Oxford University. His study of the Sykes surname obtained valid results by looking at only four markers on the male chromosome. It pointed the way to genetics becoming a valuable assistant in the service of genealogy and history.

In April 2000, Family Tree DNA began offering the first genetic genealogy tests to the public. This offering marked the first time that a personal theory on the Y chromosome could be tested outside of an academic study. Additionally, Sykes' concept of a surname study, which by this time had been adopted by several other academic researchers outside of Oxford, was expanded into online Surname Projects (an early form of social network) and the effort helped spread knowledge gained through testing to interested genealogists worldwide.

In 2001, Sykes went on to write the popular book *The Seven Daughters of Eve*, which described the seven major haplogroups of European ancestors. In the wake of the book's success, and with the growing availability and affordability of genealogical DNA testing, genetic genealogy as a field began growing rapidly. By 2003, the field of DNA testing of surnames was declared officially to have "arrived" in an article by Jobling and Tyler-Smith in *Nature Reviews Genetics*. The number of firms offering tests, and the number of consumers ordering them, had risen dramatically.

Another milestone in the acceptance of genetic genealogy is the Genographic Project. The Genographic Project is a five-year research study launched in 2005 by the National Geographic Society and IBM, in partnership with the University of Arizona and Family Tree DNA. Although its goals are primarily anthropological, not genealogical, the project's sale by April 2010 of more than 350,000 of its public participation testing kits, which test the general public for either twelve STR markers on the Y chromosome or mutations on the HVR1 region of the mtDNA, has helped increase the visibility of genetic genealogy.

More state-of-the-art commercial laboratories now recommend testing at least 25 markers, since the more markers tested, the more discriminating and powerful the results will be. A 12-marker STR test is usually not discriminating enough to provide conclusive results for a common surname. Genetic laboratories such as Genebase and Family Tree DNA give the option of testing 67 Y-DNA Markers.

Annual sales of genetic genealogical tests for all companies, including the laboratories that support them, are estimated to be in the area of $60 million (2006).

Interpretation

Since the year 2000, dozens of relevant academic papers have been published, and thousands of private test results organised by surname study groups have been made available on the internet. The comparison of results may be complicated by the fact that some laboratories use different testing methods. Apparently differing results from two sources may in fact be identical, and vice-versa.

Uses

Paternal and maternal lineages via DNA testing

The two most common types of genetic genealogy tests are Y-DNA (paternal line) and mtDNA (maternal line) genealogical DNA tests. Note that Y chromosome and Y-DNA are used interchangeably here.

These tests involve the comparison of certain sequences of the DNA of pairs of individuals in order to estimate the probability that they share a common ancestor in a genealogical time frame and, through the use of a Bayesian model published by Bruce Walsh, to estimate the number of generations separating the two individuals from their most recent common ancestor or "mrca".

Y-DNA testing involves short tandem repeat (STR) and, sometimes, single nucleotide polymorphism (SNP) testing of the Y-chromosome. The Y-chromosome is present only in males and reveals information on the strict paternal line. These tests can provide insight into the recent (via STRs) and ancient (via SNPs) genetic ancestry. A Y-chromosome STR test will reveal a haplotype, which should be similar among all male

descendants of a male ancestor. SNP tests are used to assign people to a paternal haplogroup, which defines a much larger genetic population.

mtDNA testing involves sequencing or testing the HVR-1 region, HVR-2 region or both. An mtDNA test may also include the additional SNPs needed to assign people to a maternal haplogroup—or even include the complete mtDNA.

Either Y-DNA or mtDNA test results can be compared to the results of others via private or public DNA databases.

Biogeographical and ethnic origins

Additional DNA tests exist for determining biogeographical and ethnic origin, but these tests have less relevance for traditional genealogy.

Genetic genealogy has revealed astonishing links between peoples. For instance, it has shown that the ancient Phoenician people were ancestors of much of the present-day population of the island of Malta. Preliminary results from a study by Pierre Zalloua of the American University of Beirut and Spencer Wells, supported by a grant from National Geographic's Committee for Research and Exploration, were published in the October 2004 issue of *National Geographic*. One of the conclusions is that "more than half of the Y chromosome lineages that we see in today's Maltese population could have come in with the Phoenicians."

Human migration

Genealogical DNA testing methods are also being used on a longer time scale to trace human migratory patterns. For example, they have been used to determine when the first humans came to North America and what path they followed.

For several years, a number of researchers and laboratories from around the world have been sampling indigenous populations from around the globe in an effort to map historical human migration patterns. Recently, several projects have been created that are aimed at bringing this science to the public. One example, mentioned in History above, is the National Geographic Society's Genographic Project, which aims to map historical human migration patterns by collecting and analyzing DNA samples from over 100,000 people across five continents. Another example is the DNA Clans Genetic Ancestry Analysis, which measures a person's precise genetic connections to indigenous ethnic groups from around the world.

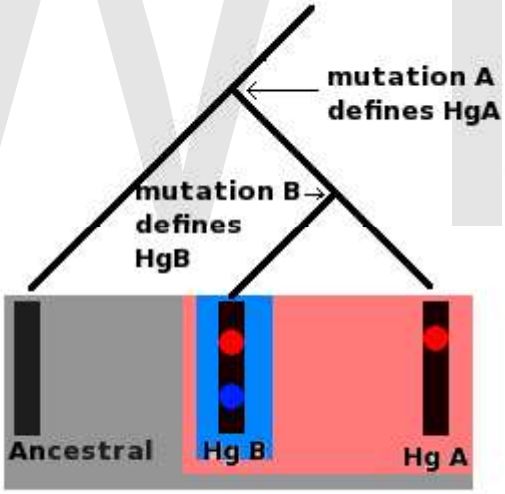
Typical customers and interest groups

Male DNA testing customers most often start with a Y chromosome test to determine their father's paternal ancestry. Females generally begin with a mitochondrial test to trace their ancient maternal lineage, which males often have tested for the same purpose.

A common consumer goal in purchasing DNA testing services is to acquire quantified, scientific linkage to a specific ancestral group. A compelling example of this motive is found in the expressed desires of some consumers to be proven to have Viking paternal ancestry. In keeping with this marketplace demand, one British DNA testing service, Oxford Ancestors, offers a Y chromosome test purporting to assess whether given males are of "Viking stock." Those whose DNA falls into the designated haplogroup are issued Viking Descendant certificates by the testing service. The same DNA testing company participated in producing a televised documentary, "The Blood of the Vikings," in conjunction with the BBC, which showed how DNA testing could reveal Viking ancestry.

The RootsWeb Genealogy-DNA Internet discussion group has a membership of 750 subscribers from around the world. Some subscribers have had various DNA tests performed and are seeking advice and guidance in interpreting their results. The list also includes administrators of DNA projects that examine surnames, geographic regions, or ethnic groups. The sophistication of subscribers ranges from expert to novice. In some cases, subscribers have been credited with making useful and novel contributions to knowledge in the field of genetic genealogy.

Paternal and maternal DNA lineages



Ancestral Haplogroup
Haplogroup A (Hg A)
Haplogroup B (Hg B)

All of these molecules are part of the ancestral haplogroup, but at some point in the past a mutation occurred in the ancestral molecule, mutation A, which produced a new lineage; this is haplogroup A and is defined by mutation A. At some more recent point in the past, a new mutation, mutation B, occurred in a person carrying haplogroup A; mutation B

defined haplogroup B. Haplogroup B is a subgroup, or subclade of haplogroup A; both haplogrups A and B are subclades of the ancestral haplogroup.

Mitochondria are small organelles that lie in the cytoplasm of eukaryotic cells, such as those of humans. Their primary purpose is to provide energy to the cell. Mitochondria are thought to be the vestigial remains of symbiotic bacteria that were once free living. One indication that mitochondria were once free living is that they contain a relatively small circular segment of DNA, called mitochondrial DNA (mtDNA). The overwhelming majority of a human's DNA is contained in chromosomes in the nucleus of the cell, but mtDNA is an exception. Individuals inherit their cytoplasm and the organelles it contains exclusively from their mothers, as these are derived from the ovum (egg cell) only, not from the sperm.

When a mutation arises in mtDNA molecule, the mutation is therefore passed in a direct female line of descent. These rare mutations are derived from copying mistakes—when the DNA is copied it is possible that a single mistake occurs in the DNA sequence, an outcome which is called a single nucleotide polymorphism (SNP).

Human Y chromosomes are male-specific sex chromosomes; nearly all humans that possess a Y chromosome will be morphologically male. Y chromosomes are therefore passed from father to son; although Y chromosomes are situated in the cell nucleus, they only recombine with the X chromosome at the ends of the Y chromosome; the vast majority of the Y chromosome (95%) does not recombine. When mutations (SNPs, and STR copying mistakes) arise in the Y chromosome, they are passed down directly from father to son in a direct male line of descent. The Y-DNA and mtDNA therefore share a certain feature: they both pass down unchanged except for mutations.

The other chromosomes, autosomes and X chromosomes in women, share their genetic material (called crossing over leading to recombination) during meiosis (a special type of cell division that occurs for the purposes of sexual reproduction). Effectively this means that the genetic material from these chromosomes gets mixed up in every generation, and so any new mutations are passed down randomly from parents to offspring.

The special feature that both Y-DNA and mtDNA share, above, preserves a "written" record of their mutations because neither DNA gets mixed up or randomized—mutations remain fixed in place on both types of DNA. Furthermore the historical sequence of these mutations can also be inferred. For example, if a set of ten Y chromosomes (derived from ten different men) contains a mutation, A, but only five of these chromosomes contain a second mutation, B, it must be the case that mutation B occurred after mutation A.

Furthermore all ten men who carry the chromosome with mutation A are the direct male line descendants of the same man who was the first to carry this mutation. The first man to carry mutation B was also a direct male line descendant of this man, but is also the direct male line ancestor of all men carrying mutation B. Series of mutations such as this form molecular lineages. Furthermore each SNP mutation may define a set of specific Y chromosomes called a haplogroup.

All men carrying SNP mutation A form a single haplogroup, and all men carrying mutation B are part of this haplogroup, but mutation B (if a SNP) may also define a more recent haplogroup (which is a subgroup or subclade) of its own which men carrying only mutation A do not belong to. Both mtDNA and Y chromosomes or Y-DNA are grouped into lineages and haplogroups; these are often presented as tree-like diagrams.

Benefits

Genetic genealogy gives genealogists a means to check or supplement their genealogy results with information obtained via DNA testing. A positive test match with another individual may:

- provide locations for further genealogical research
- help determine ancestral homeland
- discover living relatives
- validate existing research
- confirm or deny suspected connections between families
- prove or disprove theories regarding ancestry

Drawbacks

People who resist testing may cite one of the following concerns:

- Cost
- Quality of testing
- Concerns over privacy issues
- Loss of ethnic identity

Finally, Y-DNA and mtDNA tests each only trace a single lineage (one's father's father's father's etc. lineage or one's mother's mother's mother's etc. lineage). At 10 generations back, an individual has up to 1024 unique ancestors (fewer if ancestor cousins interbred) and a Y-DNA or mtDNA test is only studying one of those ancestors, as well as their descendants and siblings (same sexed siblings for Y-DNA or all siblings for mtDNA). However, most genealogists maintain contact with many cousins (1st, 2nd, 3rd, etc., with different surnames) whose Y-DNA and mtDNA are different, and thus can be encouraged to be tested to find additional ancestral DNA lineages.

Expected growth

Genetic genealogy is a rapidly growing field. As the cost of testing continues to drop, the number of people being tested continues to increase. The probability of finding a genetic match among the DNA databases should continue to improve. Laboratories and testing firms are engaging in active research and development that will allow for higher confidence intervals and better results interpretation, including historical interpretive reports and customized research.

Genetic distance among relatives

Where the genogram or family tree of individuals is known, it can be used to determine the genetic identity between individuals. It is often described as percentage of genetic identity, referring to the fraction of genome inherited from common ancestors, and not actual genomic identity, which is always approximately 99.9% identical from one human to another.

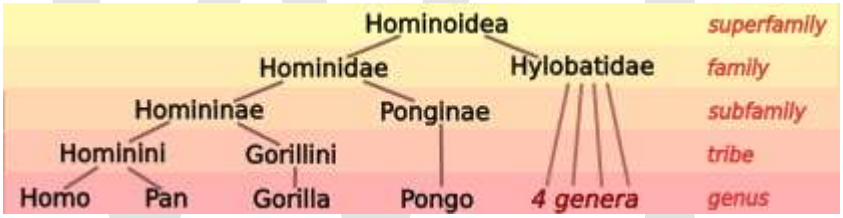
One method of calculating this genetic similarity is to do an inbreeding calculation by the path or tabular method and then multiply by 2, because any progeny would have a 1 in 2 risk of actually inheriting the identical alleles from both parents. For instance, a brother/sister relation gives 25% risk for two alleles to be identical by descent.

Chapter- 3

Human Evolutionary Genetics

Human evolutionary genetics studies how one human genome differs from the other, the evolutionary past that gave rise to it, and its current effects. Differences between genomes have anthropological, medical and forensic implications and applications. Genetic data can provide important insight into human evolution.

Origin of apes

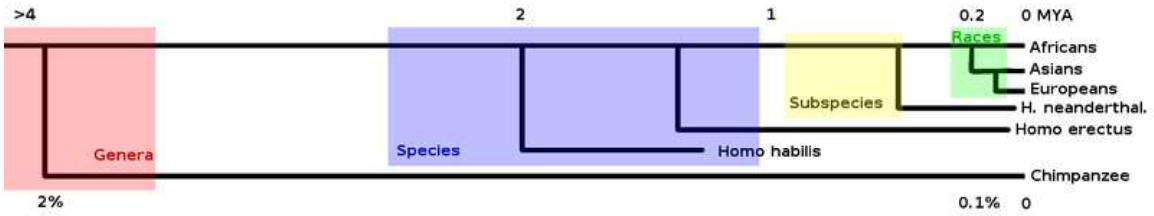


Taxonomic relationships of hominoids

Biologists classify humans, along with only a few other species, as great apes (species in the family Hominidae). The Hominidae include two distinct species of chimpanzee (the bonobo, *Pan paniscus*, and the common chimpanzee, *Pan troglodytes*), two species of gorilla (the western gorilla, *Gorilla gorilla*, and the eastern gorilla, *Gorilla graueri*), and two species of orangutan (the Bornean orangutan, *Pongo pygmaeus*, and the Sumatran orangutan, *Pongo abelii*).

Apes, in turn, belong to the primates order (>400 species). Data from both mitochondrial DNA (mtDNA) and nuclear DNA (nDNA) indicates that primates belong to the group of Euarchontoglires, together with Rodentia, Lagomorpha, Dermoptera, and Scandentia. This is further supported by Alu-like short interspersed nuclear elements (SINEs) which have been found only in members of the Euarchontoglires.

Cladistics



A phylogenetic tree like the one shown above is usually derived from DNA or protein sequences from populations. Often mitochondrial DNA or Y chromosome sequences are used to study ancient human demographics. These single-locus sources of DNA do not recombine and are almost always inherited from a single parent, with only one known exception in mtDNA (Schwartz and Vissing 2002). Individuals from the various continental groups tend to be more similar to one another than to people from other continents. The tree is rooted in the common ancestor of chimpanzees and humans, which is believed to have originated in Africa. Horizontal distance in the diagram corresponds to two things:

1. **Genetic distance**. Given below the diagram, the genetic difference between humans and chimps is less than 2%, or 20 times larger than the variation among modern humans.
2. **Temporal remoteness** of the most recent common ancestor. Rough estimates are given above the diagram, in millions of years. The mitochondrial most recent common ancestor of modern humans lived roughly 200,000 years ago, latest common ancestors of humans and chimps between four and seven million years ago.

Chimpanzees and humans belong to different genera, indicated in red. Formation of species and subspecies is also indicated, and the formation of "races" is indicated in the green rectangle to the right (note that only a very rough representation of human phylogeny is given). Note that vertical distances are not meaningful in this representation.

Speciation of humans and the African apes

The separation of humans from their closest relatives, the African apes (chimpanzees and gorillas), has been studied extensively for more than a century. Five major questions have been addressed:

- Which apes are our closest ancestors?
- When did the separations occur?
- What was the effective population size of the common ancestor before the split?
- Are there traces of population structure (subpopulations) preceding the speciation or partial admixture succeeding it?
- What were the specific events (including fusion of chromosomes 2a and 2b) prior to and subsequent to the separation?

General observations

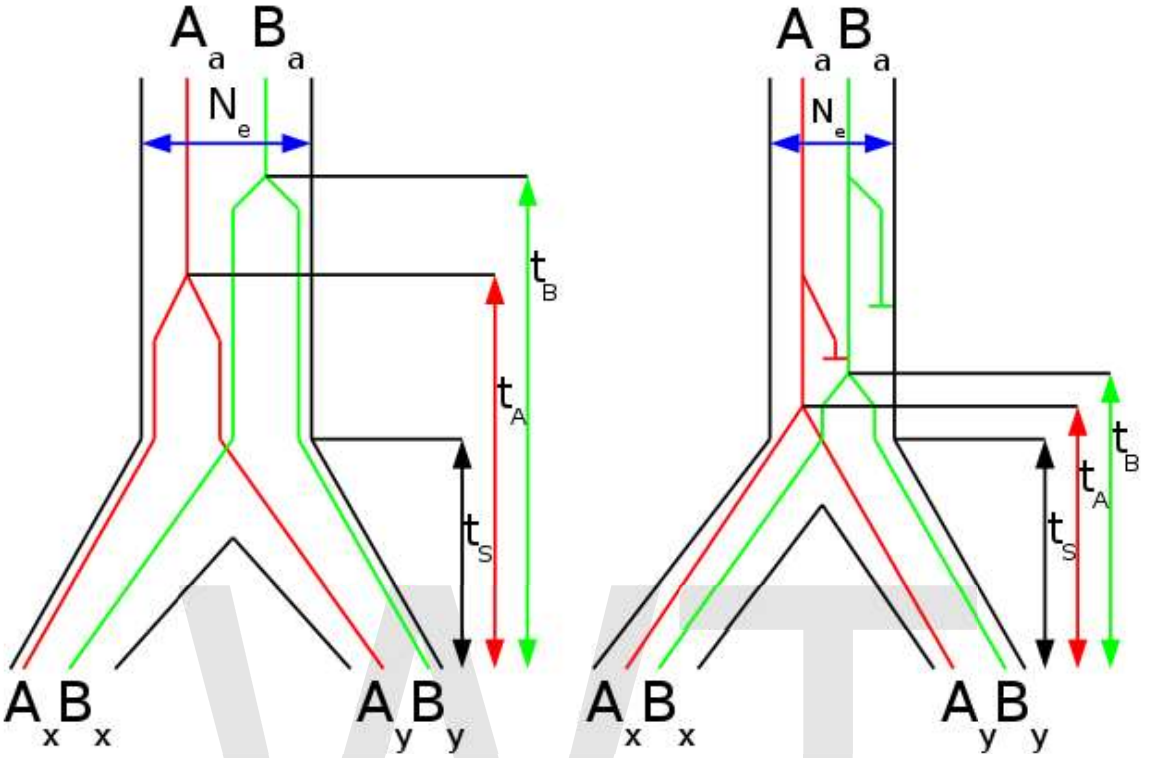
As discussed before, different parts of the genome show different sequence divergence between different hominoids. It has also been shown that the sequence divergence between DNA from humans and chimpanzees varies greatly. For example the sequence divergence varies between 0% to 2.66% between non-coding, non-repetitive genomic regions of humans and chimpanzees. Additionally gene trees, generated by comparative analysis of DNA segments, do not always fit the species tree. Summing up:

- The sequence divergence varies significantly between humans, chimpanzees and gorillas.
- For most DNA sequences, humans and chimpanzees appear to be most closely related, but some point to a human-gorilla or chimpanzee-gorilla clade.
- The human genome has been sequenced, as well as the chimpanzee genome. Humans have 23 pairs of chromosomes, while chimpanzees, gorillas, and orangutans have 24. Human chromosome 2 is a fusion between two chromosomes that remained separate in the other primates.

Divergence times

The divergence time of humans from other apes is of great interest. One of the first molecular studies, published in 1967 measured immunological distances (IDs) between different primates. Basically the study measured the strength of immunological response that an antigen from one species (human albumin) induces in the immune system of another species (human, chimpanzee, gorilla and Old World monkeys). Closely related species should have similar antigens and therefore weaker immunological response to each other's antigens. The immunological response of a species to its own antigens (e.g. human to human) was set to be 1. The ID between humans and gorillas was determined to be 1.09, that between humans and chimpanzees was determined as 1.14. However the distance to six different Old World monkeys was on average 2.46 indicating that the African apes are far closer related to humans than to monkeys. The authors consider the divergence time between Old World monkeys and hominoids to be 30 million years ago (MYA), based on fossil data, and the immunological distance was considered to grow at a constant rate. They concluded that divergence time of humans and the African apes to be roughly ~5 MYA. That was a surprising result. Most scientists at that time thought that humans and great apes diverged much earlier (>15 MYA). The gorilla was, in ID terms, closer to human than to chimpanzees, however the difference was so slight that the trichotomy could not be resolved with certainty. Later studies based on molecular genetics were able to resolve the trichotomy: chimpanzees are phylogenetically closer to humans than to gorillas. However, the divergence times estimated later (using much more sophisticated methods in molecular genetics) do not substantially differ from the very first estimate in 1967.

Divergence times and ancestral effective population size

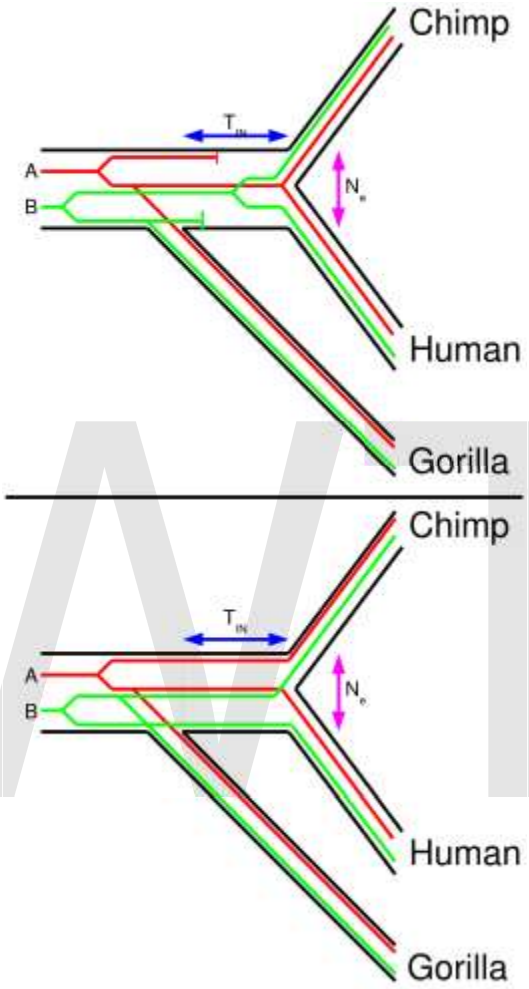


The sequences of the DNA segments diverge earlier than the species. A large effective population size in the ancestral population (left) preserves different variants of the DNA segments (=alleles) for a longer period of time. Therefore, on average, the gene divergence times (t_A for DNA segment A; t_B for DNA segment B) will deviate more from the time the species diverge (t_S) compared to a small ancestral effective population size (right).

Current methods to determine divergence times use DNA sequence alignments and molecular clocks. Usually the molecular clock is calibrated assuming that the orangutan split from the African apes (including humans) 12-16 MYA. Some studies also include some old world monkeys and set the divergence time of them from hominoids to 25-30 MYA. Both calibration points are based on very little fossil data and have been criticized. If these dates are revised, the divergence times estimated from molecular data will change as well. However, the relative divergence times are unlikely to change. Even if we can't tell absolute divergence times exactly, we can be pretty sure that the divergence time between chimpanzees and humans is about sixfold shorter than between chimpanzees (or humans) and monkeys.

One study (Takahata *et al.*, 1995) used 15 DNA sequence from different regions of the genome from human and chimpanzee and 7 DNA sequences from human, chimpanzee and gorilla. They determined that chimpanzees are more closely related to humans than gorillas. Using various statistical methods, they estimated the divergence time human-chimp to be 4.7 MYA and the divergence time between gorillas and humans (and

chimps) to be 7.2 MYA. Additionally they estimated the effective population size of the common ancestor of humans and chimpanzees to be ~100,000. This was somewhat surprising since the present day effective population size of humans is estimate to be only ~10,000. If true that means that the human lineage would have experienced an immense decrease of its effective population size (and thus genetic diversity) in its evolution.



A and B are two different loci. In the upper figure they fit to the species tree. The DNA that is present in today's gorillas diverged earlier from the DNA that is present in today's humans and chimps. Thus both loci should be more similar between human and chimp than between gorilla and chimp or gorilla and human. In the lower graph, locus A has a more recent common ancestor in human and gorilla compared to the chimp sequence. Whereas chimp and gorilla have a more recent common ancestor for locus B. Here the gene trees are incongruent to the species tree.

Another study (Chen & Li, 2001) sequenced 53 non-repetitive, intergenic DNA segments from a human, a chimpanzee, a gorilla, and orangutan. When the DNA sequences were concatenated to a single long sequence, the generated neighbor-joining tree supported the *Homo-Pan* clade with 100% bootstrap (that is that humans and chimpanzees are the

closest related species of the four). When three species are fairly closely related to each other (like human, chimpanzee and gorilla), the trees obtained from DNA sequence data may not be congruent with the tree that represents the speciation (species tree). The shorter internodal time span (T_{IN}) the more common are incongruent gene trees. The effective population size (N_e) of the internodal population determines how long genetic lineages are preserved in the population. A higher effective population size causes more incongruent gene trees. Therefore, if the internodal time span is known, the ancestral effective population size of the common ancestor of humans and chimpanzees can be calculated.

When each segment was analyzed individually, 31 supported the *Homo*-*Pan* clade, 10 supported the *Homo*-*Gorilla* clade, and 12 supported the *Pan*-*Gorilla* clade. Using the molecular clock the authors estimated that gorillas split up first 6.2-8.4 MYA and chimpanzees and humans split up 1.6-2.2 million years later (internodal time span) 4.6-6.2 MYA. The internodal time span is useful to estimate the ancestral effective population size of the common ancestor of humans and chimpanzees.

A parsimonious analysis revealed that 24 loci supported the *Homo*-*Pan* clade, 7 supported the *Homo*-*Gorilla* clade, 2 supported the *Pan*-*Gorilla* clade and 20 gave no resolution. Additionally they took 35 protein coding loci from databases. Of these 12 supported the *Homo*-*Pan* clade, 3 the *Homo*-*Gorilla* clade, 4 the *Pan*-*Gorilla* clade and 16 gave no resolution. Therefore only ~70% of the 52 loci that gave a resolution (33 intergenic, 19 protein coding) support the 'correct' species tree. From the fraction of loci which did not support the species tree and the internodal time span they estimated previously, the effective population of the common ancestor of humans and chimpanzees was estimated to be ~52 000 to 96 000. This value is not as high as that from the first study (Takahata), but still much higher than present day effective population size of humans.

A third study (Yang, 2002) used the same dataset that Chen and Li used but estimated the ancestral effective population of 'only' ~12,000 to 21,000, using a different statistical method.

Genetic differences between humans and other great apes

The alignable sequences within genomes of humans and chimpanzees differ by about 35 million single nucleotide substitutions. Additionally about 3% of the complete genomes differ by deletions, insertions and duplications.

Since mutation rate is relatively constant, roughly one half of these changes occurred in the human lineage. Only a very tiny fraction of those fixed differences gave rise to the different phenotypes of humans and chimpanzees and finding those is a great challenge. The vast majority of the differences are neutral and do not affect the phenotype.

Molecular evolution may act in different ways, through protein evolution, gene loss, differential gene regulation and RNA evolution. All are thought to have played some part in human evolution.

Gene loss

Many different mutations can inactivate a gene, but few will change its function in a specific way. Inactivation mutations will therefore be readily available for selection to act on. Gene loss could thus be a common mechanism of evolutionary adaptation (the "less-is-more" hypothesis).

80 genes were lost in the human lineage after separation from the last common ancestor with the chimpanzee. 36 of those were for olfactory receptors. Genes involved in chemoreception and immune response are overrepresented. Another study estimated that 86 genes had been lost.

Hair keratin gene KRTHAP1

A gene for type I hair keratin was lost in the human lineage. Keratins are a major component of hairs. Humans still have nine functional type I hair keratin genes but the loss of that particular gene may have caused the thinning of human body hair. The gene loss occurred relatively recently in human evolution—less than 240,000 years ago.

Myosin gene MYH16

Stedman *et al.* (2004) stated that the loss of the sarcomeric myosin gene MYH16 in the human lineage led to smaller masticatory muscles. They estimated that the mutation that led to the inactivation (a two base pair deletion) occurred 2.4 million years ago, predating the appearance of *Homo ergaster/erectus* in Africa. The period that followed was marked by a strong increase in cranial capacity, promoting speculation that the loss of the gene may have removed an evolutionary constraint on brain size in the genus *Homo*.

Another estimate for the loss of the MYH16 gene is 5.3 million years ago, long before *Homo* appeared.

Other

- CASPASE12, a cysteinyl aspartate proteinase

Gene addition

Segmental duplications (SDs or LCRs) have had roles in creating new primate genes and shaping human genetic variation.

Selection pressures

Human accelerated regions are areas of the genome that differ between humans and chimpanzees to a greater extent than can be explained by genetic drift over the time since the two species shared a common ancestor. These regions show signs of being subject to natural selection, leading to the evolution of distinctly human traits. Two examples are HAR1F, which is believed to be related to brain development and HAR2 (a.k.a HACNS1) that may have played a role in the development of the opposable thumb.

Genetic differences between humans and Neanderthals

An international group of scientists completed a draft sequence of the Neanderthal genome in May 2010. The results indicate some breeding between humans and Neanderthals as the genomes of non-African humans have 1-4% more in common with Neanderthals than do the genomes of subsaharan Africans. Neanderthals and most humans share a lactose-intolerant variant of the lactase gene that encodes an enzyme that is unable to break down lactose in milk after weaning. Humans and Neanderthals also share the FOXP2 gene variant associated with brain development and with speech in humans, indicating that Neanderthals may have been able to speak. Chimps have two amino acid differences in FOXP2 compared with human and Neanderthal FOXP2.

Sequence divergence between humans and apes

The draft sequence of the common chimpanzee genome published in the summer 2005 showed the regions that are similar enough to be aligned with one another account for 2400 million of the human genome's 3164.7 million bases – that is, 75.8% of the genome. This 75.8% of the human genome is 1.23% different from the chimpanzee genome in single nucleotide polymorphisms (changes of single DNA "letters" in the genome). Another type of difference, called indels (insertions/deletions) account for another ~3 % difference between the alignable sequences. In addition, variation in copy number of large segments (> 20 kb) of similar DNA sequence provides a further 2.7% difference between the two species. Hence the total similarity of the genomes could be as low as about 70%.

The figures above do not take into account differences in the organization of the alignable sequences within the genomes of humans and chimps. Short stretches of alignable sequence may be in very different orders and locations within the two genomes. At present we cannot fully assess the difference in structure of the two genomes, because the human genome was used as a scaffold when the chimpanzee draft genome was assembled. When genomes are sequenced, relatively short sequences of DNA are produced, and these sequences have to be fitted together like a jigsaw puzzle. This requires multiple overlapping reads to accurately assemble the overall sequence. The human genome sequence is relatively accurate, with 8 to 9-fold coverage, but the chimpanzee draft genome only has 3.6-fold coverage. The human genome was sequenced using a hierarchical shotgun method which can deal with duplications and difficult-to-assemble sequences better than the whole genome shotgun method that was used for the

chimpanzee draft genome. The human genome was used as a template for the assembly of the draft chimpanzee genome, on the assumption that the two genomes would be similar.

Almost half of that 1.23% SNP change belongs to the human at 0.53%, whose genetic variance is lower than a chimp, and just over half to the chimp at 0.7%. If we also take into account that random "genetic drift" takes up the bulk of the 0.54% difference, then that percentage difference where SNPs have a potential positive impact on human abilities, is between 0.01% and 0.02%. The bonobo is a sibling species of common chimpanzee and is genetically about as different from humans as are common chimps.

Percentage sequence divergence between humans and other hominids

Locus	Human-Chimp	Human-Gorilla	Human-Orangutan
Alu elements	2	-	-
Non-coding (Chr. Y)	1.68 ± 0.19	2.33 ± 0.2	5.63 ± 0.35
Pseudogenes (autosomal)	1.64 ± 0.10	1.87 ± 0.11	-
Pseudogenes (Chr. X)	1.47 ± 0.17	-	-
Noncoding (autosomal)	1.24 ± 0.07	1.62 ± 0.08	3.08 ± 0.11
Genes (K_s)	1.11	1.48	2.98
Introns	0.93 ± 0.08	1.23 ± 0.09	-
Xq13.3	0.92 ± 0.10	1.42 ± 0.12	3.00 ± 0.18
Subtotal for X chromosome	1.16 ± 0.07	1.47 ± 0.08	-
Genes (K_a)	0.8	0.93	1.96

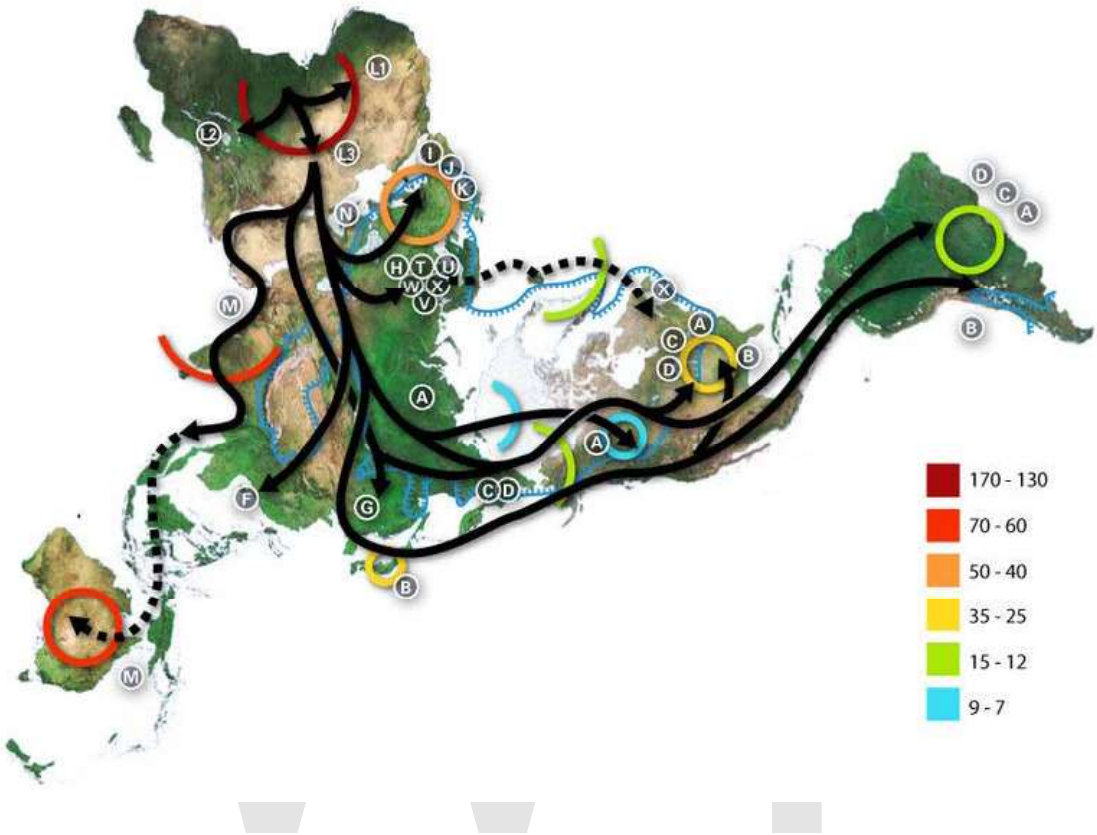
The sequence divergence has generally the following pattern: Human-Chimp < Human-Gorilla << Human-Orangutan, highlighting the close kinship between humans and the African apes. Alu elements diverge quickly due to their high frequency of CpG dinucleotides which mutate roughly 10 times more often than the average nucleotide in the genome. The mutation rate is higher in the male germ line, therefore the divergence in the Y chromosome—which is inherited solely from the father—is higher than in autosomes. The X chromosome is inherited twice as often through the female germ line as through the male germ line and therefore shows slightly lower sequence divergence. The sequence divergence of the Xq13.3 region is surprisingly low between humans and chimpanzees.

Mutations altering the amino acid sequence of proteins (K_a) are the least common. In fact ~29% of all orthologous proteins are identical between human and chimpanzee. The typical protein differs by only two amino acids.

The measures of sequence divergence shown in the table only take the substitutional differences, for example from an A (adenine) to a G (guanine), into account. DNA sequences may however also differ by insertions and deletions (indels) of bases. These are usually stripped from the alignments before the calculation of sequence divergence is

performed. The overall sequence divergence between humans and chimpanzees for example is close to 5% if indels would be included.

Modern humans



Map of the migration of modern humans out of Africa, based on mitochondrial DNA. Coloured rings indicate years before present, in thousands.

Molecular biologists starting with Wesley Brown on mtDNA and Allan Wilson on mtDNA have produced observations relevant to human evolution.

Age of the common ancestor

By estimating the rate at which mutations occur in mtDNA, the age of the common ancestral mtDNA type can be estimated: "the common ancestral mtDNA (type a) links mtDNA types that have diverged by an average of nearly 0.57%. Assuming a rate of 2%-4% per million years, this implies that the common ancestor of all surviving mtDNA types existed 140,000-290,000 years ago." This observation is robust, and this common direct female line ancestor (or mitochondrial most recent common ancestor (mtMRCA)) of all extant humans has become known as Mitochondrial Eve. The observation that the mtMRCA is the direct matrilineal ancestor of all living humans does not mean either that she was the first anatomically modern human, nor that no other female humans lived concurrently with her. Other women would have lived at the same time and passed

nuclear genes down to living humans, but their mitochondrial lineages were lost over time. This could be due to random events such as producing only male children.

African origin for modern humans

There is evidence that modern human mtDNA has an African origin: "We infer from the tree of minimum length... that Africa is a likely source of the human mitochondrial gene pool. This inference comes from the observation that one of the two primary branches leads exclusively to African mtDNAs... while the second primary branch also leads to African mtDNAs... By postulating that the common ancestral mtDNA... was African, we minimize the number of intercontinental migrations needed to account for the geographic distribution of mtDNA types."

The broad study of African genetic diversity headed by Sarah Tishkoff found the San people to express the greatest genetic diversity among the 113 distinct populations sampled, making them one of 14 "ancestral population clusters". The research also located the origin of modern human migration in south-western Africa, near the coastal border of Namibia and Angola.

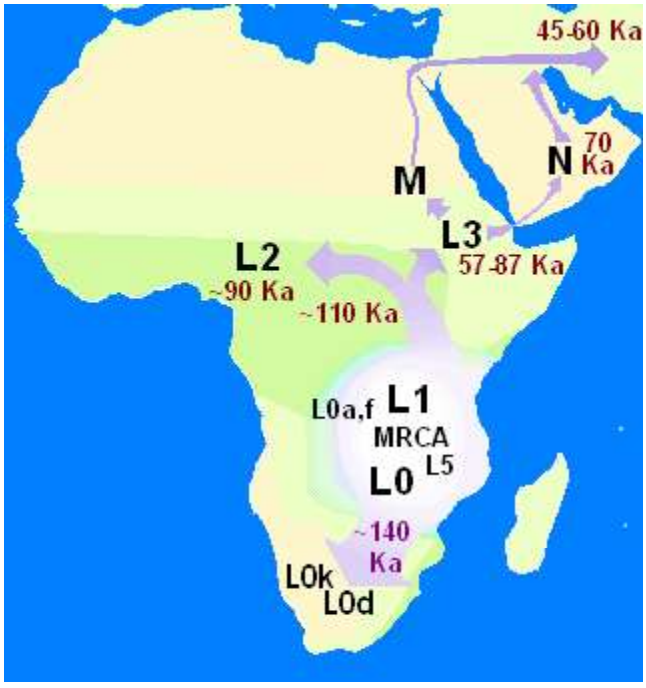
Y chromosome findings

The Y chromosome is much larger than mtDNA, and is relatively homogeneous; therefore it has taken much longer to find distinct lineages and to analyse them. Conversely, because the Y chromosome is so large by comparison, it holds more genetic information. Y chromosome studies show similar findings to those made with mtDNA. The estimate for the age of the ancestral Y chromosome for all extant Y chromosomes is given at about 70,000 years ago and is also placed in Africa; the individual who contributed this Y chromosomal heritage is sometimes referred to as Y chromosome Adam. The difference in dates between Y chromosome Adam and mitochondrial Eve is usually attributed to a higher extinction rate for Y chromosomes due to greater differential reproductive success between individual men, which means that a small number of very successful men may produce many children, while a larger number of less successful men will produce far fewer children.

Chapter- 4

Mitochondrial Eve

Haplogroup Modern humans



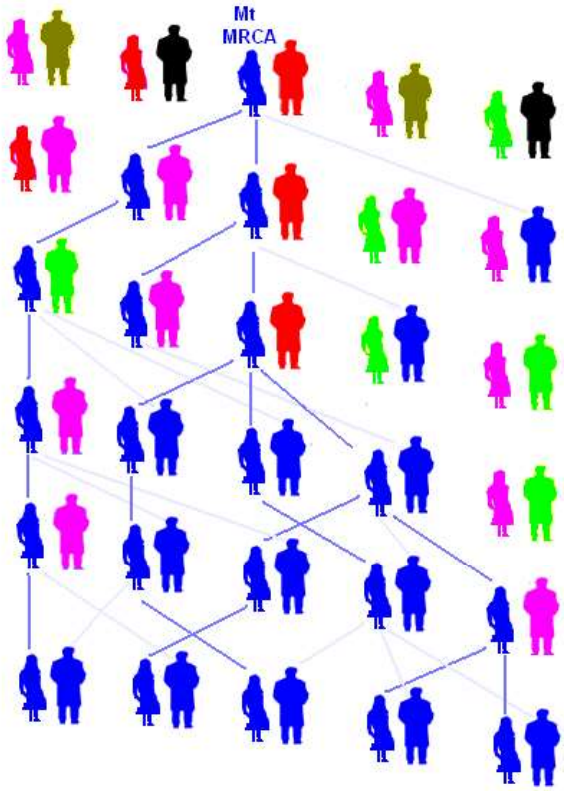
Possible time of origin	152,000 - 234,000 BP
Possible place of origin	East Africa
Ancestor	n/a
Descendants	Mitochondrial macro-haplogroups L0, L1, and L5
Defining mutations	None

In the field of human genetics, **Mitochondrial Eve** refers to the matrilineal "MRCA" (most recent common ancestor). In other words, this was the woman from whom all living humans today descend, on their mother's side, and through the mothers of those mothers and so on, back until all lines converge on one person. Because it is passed from mother to offspring without recombination, all mitochondrial DNA (mtDNA) in every living person is directly descended from hers by definition. Mitochondrial Eve is the female counterpart of Y-chromosomal Adam, the patrilineal most recent common ancestor, although they lived thousands of years apart.

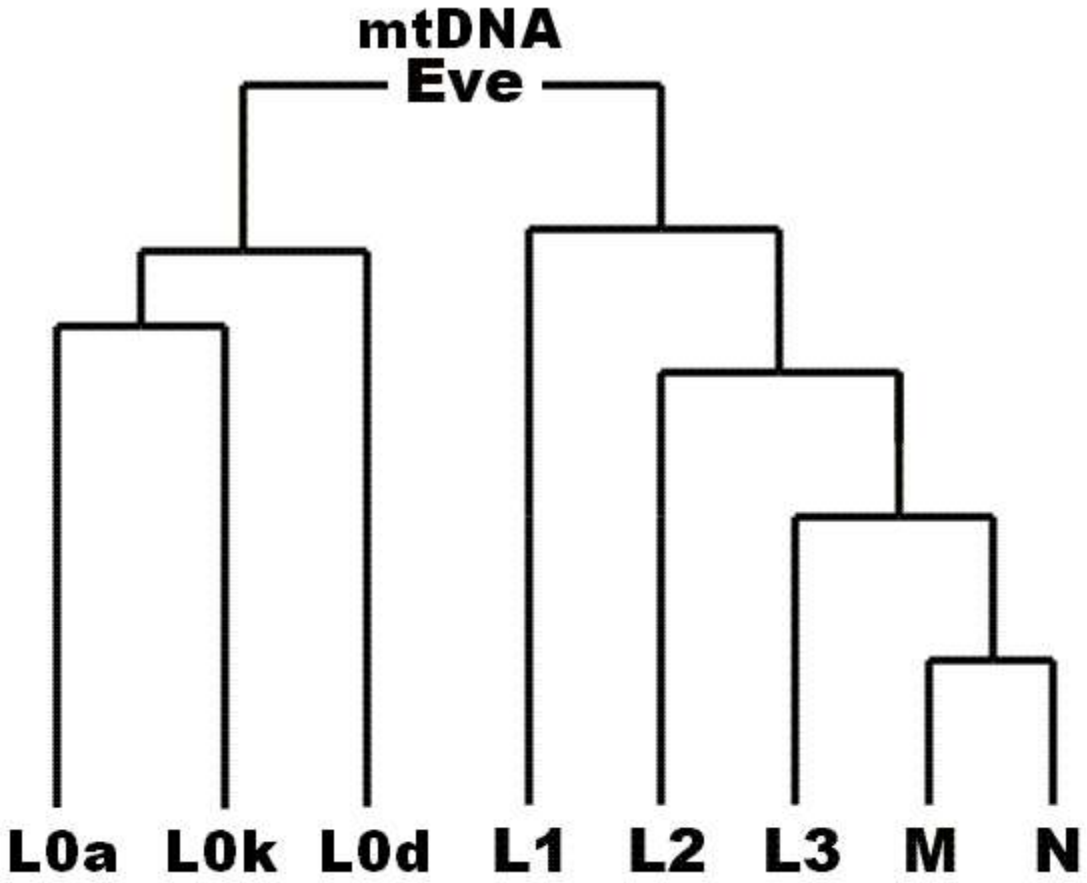
Mitochondrial Eve is generally estimated to have lived around 200,000 years ago, most likely in East Africa, when *Homo sapiens sapiens* ("anatomically modern humans") were developing as a population distinct from other human sub-species.

Mitochondrial Eve lived much earlier than the out of Africa migration that is thought to have occurred between 95,000 to 45,000 BP. The dating for 'Eve' was a blow to the multiregional hypothesis, and a boost to the hypothesis that modern humans originated relatively recently in Africa and spread from there, replacing more "archaic" human populations such as Neanderthals. As a result, the latter hypothesis is now the dominant one.

Female and mitochondrial ancestry



Through random drift or selection the female-lineage may trace back to a single female, such as Mitochondrial Eve



Simplified Human mitochondrial phylogeny

Without a DNA sample, it is not possible to reconstruct the complete genetic makeup (genome) of any individual who died very long ago. By analysing descendants' DNA, however, parts of ancestral genomes are estimated by scientists. Mitochondrial DNA (mtDNA) and Y chromosome are commonly used to trace ancestry in this manner. mtDNA is generally passed un-mixed from mothers to children of both sexes, along the maternal line, or matrilineally. Matrilineal descent goes back to our mothers, to their mothers, until all female lineages converge.

Branches are identified by one or more unique markers which give a mitochondrial "DNA signature" or "haplotype" (e.g. the CRS is a haplotype). Each marker is a DNA base-pair that has resulted from an SNP mutation. Scientists sort mitochondrial DNA results into more or less related groups, with more or less recent common ancestors. This leads to the construction of a DNA family tree where the branches are in biological terms clades, and the common ancestors such as Mitochondrial Eve sit at branching points in this tree. Major branches are said to define a haplogroup (e.g. CRS belongs to haplogroup H), and large branches containing several haplogroups are called "macro-haplogroups".

The mitochondrial clade which Mitochondrial Eve defines is the species *Homo sapiens sapiens* itself, or at least the current population or "chronospecies" as it exists today. In

principle, earlier Eves can also be defined going beyond the species, for example one who is ancestral to both modern humanity and Neanderthals, or, further back, an "Eve" ancestral to all members of genus *Homo* and chimpanzees in genus *Pan*. According to current nomenclature, Mitochondrial Eve's haplogroup was within mitochondrial haplogroup L because this macro-haplogroup contains all surviving human mitochondrial lineages today.

The variation of mitochondrial DNA between different people can be used to estimate the time back to a common ancestor, such as Mitochondrial Eve. This works because, along any particular line of descent, mitochondrial DNA accumulates mutations at the rate of approximately one every 3,500 years. A certain number of these new variants will survive into modern times and be identifiable as distinct lineages. At the same time some branches, including even very old ones, come to an end, when the last family in a distinct branch has no daughters.

Mitochondrial Eve is the most recent common matrilineal ancestor for all modern humans. Whenever one of the two most ancient branch lines dies out, the MRCA will move to a more recent female ancestor, always the most recent mother to have more than one daughter with living maternal line descendants alive today. The number of mutations that can be found distinguishing modern people is determined by two criteria: firstly and most obviously, the time back to her, but secondly and less obviously by the varying rates at which new branches have come into existence and old branches have become extinct. By looking at the number of mutations which have been accumulated in different branches of this family tree, and looking at which geographical regions have the widest range of least related branches, the region where Eve lived can be proposed.

The date when Mitochondrial Eve lived is estimated by determining the MRCA of a sample of mtDNA lineages. In 1980, Brown first proposed that modern humans possessed a mitochondrial common ancestor that may have lived as recently as 180 kya. In 1987, Cann et al. suggested that mitochondrial Eve may have lived between 140-280 kya.

Common fallacies

Not the only woman

One of the misconceptions of mitochondrial Eve is that since all women alive today descended in a direct unbroken female line from her that she was the only woman alive at the time. However nuclear DNA studies indicate that the size of the ancient human population never dropped below some tens of thousands; there were many other women around at Eve's time with descendants alive today, but somewhere in all *their* lines of descent there is at least one generation with no female offspring (and men do not pass on their mothers' mitochondrial DNA to their children). By contrast, Eve's lines of descent to each person alive today includes at least one line of descent to each person which is purely matrilineal.

Not a contemporary of "Adam"

Sometimes mitochondrial Eve is assumed to have lived at the same time as Y-chromosomal Adam, perhaps even meeting and mating with him. However there is no such parallel with the Biblical story. Like mitochondrial "Eve", Y-chromosomal "Adam" probably lived in Africa; however, this "Eve" lived much earlier than this "Adam" – perhaps some 50,000 to 80,000 years earlier.

Not the most recent ancestor shared by all humans

Mitochondrial Eve is the most recent common *matrilineal* ancestor, not the *most recent common ancestor* (MRCA). Since the mtDNA is inherited maternally and recombination is either rare or absent, it is relatively easy to track the ancestry of the lineages back to a MRCA; however this MRCA is valid only when discussing mitochondrial DNA. An approximate sequence from newest to oldest can list various important points in the ancestry of modern human populations:

- The Human MRCA. All humans alive today share a surprisingly recent common ancestor, perhaps even within the last 5,000 years, even for people born on different continents.
- The Identical ancestors point. Just a few thousand years before the most recent single ancestor shared by all living humans comes the time at which all humans who were alive either left no descendants or are common ancestors to all humans alive today. In other words, from this point back in time "each present-day human has exactly the same set of genealogical ancestors". This is far more recent than Mitochondrial Eve.
- "Y-Chromosomal Adam", the most recent male-line ancestor of all living men, was much more recent than Mitochondrial Eve, but is also likely to have been long before the Identical ancestors point.

Implications of dating and placement of Eve

Initially there was a lot of resistance against the Mitochondrial Eve hypothesis. This resistance was due, in part, to the popularity of the *Multiregional Evolution hypothesis* amongst some leading paleoanthropologists such as Milford Wolpoff. This prevailing theory held that the evolution of humanity from the beginning of the Pleistocene 2.5 million years BP to the present day has been within a single, continuous human species, evolving worldwide to modern Homo sapiens. More resistance came from those who argued that there was too little time between *Homo erectus* and modern *Homo sapiens* to allow for another new species, and others who argued that for regional evolution from archaic hominin forms into modern ones. Consequently, the finding of a recent maternal ancestor for all humans in Africa was very controversial.

Cann, Stoneking & Wilson (1987)'s placement of a relatively small population of humans in sub-saharan Africa, lent appreciable support for the recent *Out of Africa* hypothesis. The current concept places between 1,500 and 16,000 effectively interbreeding

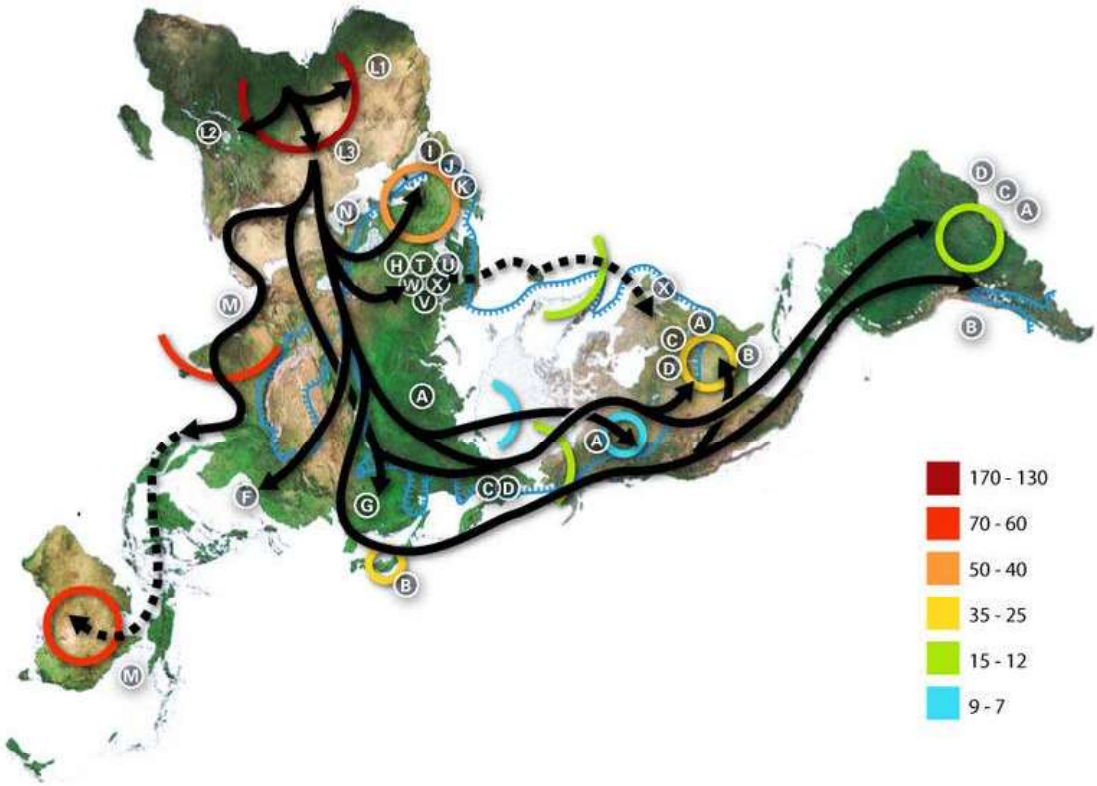
individuals (census 4,500 to 48,000 individuals) within Tanzania and proximal regions. Later, Tishkoff using data from many loci has extrapolated origins to the Angola-Namibia border region near the Atlantic Ocean (although this region has poor genetic definition), whereas Behar et al. 2008 places an ancestral population in Ethiopia. These opinions all point toward a sub-Saharan origin. More recent literature on languages and pygmy phenotype indicate that L0 and L1 were carried by click-speaking pygmies from SE Africa to Central and Western Africa, therefore explaining much of the genetic diversity in those regions. Consequently, more recent studies have tended to push the cradle of humanity more toward the South or East of Africa.

To some extent the studies have already revealed that the presence of archaic homo sapiens in Northwest Africa (Jebel Irhoud) were not likely part of the contiguous modern human population. In addition, the older remains at Skhul and Qafzeh are also unlikely part of the constrict human population, evidence currently indicates humans expanded in the region no earlier than 90,000 BP. Tishkoff argues that humans might have migrated to the Levant before 90 Ka, but this colony did not persist in SW Asia. Better defined is the genetic separation among Neanderthals, Flores hobbit, Java man, and Peking man. In 1999 Krings et al., eliminated problems in molecular clocking postulated by Nei, 1992 when it was found the mtDNA sequence for the same region was substantially different from the MRCA relative to any human sequence. Currently there are 6 fully sequenced Neanderthal mitogenomes, each falling within a genetic cluster less diverse than that for humans, and mitogenome analysis in humans has statistically markedly reduced the TMRCA range so that it no longer overlaps with Neandertal/human split times. Of all the non-African hominids European archaics most closely resembled humans, indicating a wider genetic divide with other hominids.

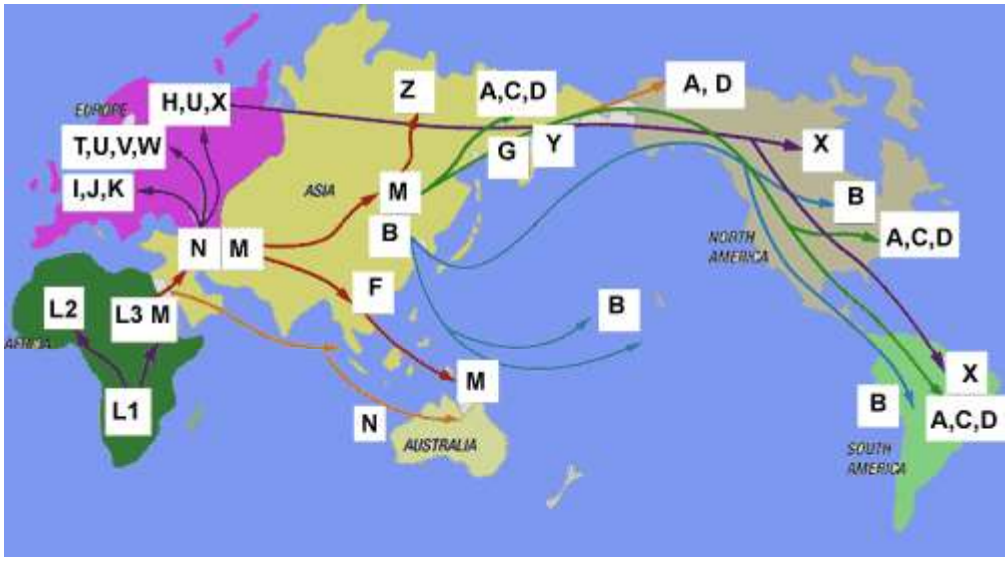
Since the multiregional evolution hypothesis (MREH) revolved around a belief that regional modern human populations evolved in situ in various regions (Europe: Neandertals to Europeans, Asia: Homo erectus to East Asians, Australia: Sumatran erectines to indigenous Australians), these results demonstrated that a pure MREH hypothesis could not explain one important genetic marker.

Chapter- 5

Human Mitochondrial DNA Haplogroup



Hypothesized map of human migration based on mitochondrial DNA.



Another model of human migration based on Mitochondrial DNA

In human genetics, a **human mitochondrial DNA haplogroup** is a haplogroup defined by differences in human mitochondrial DNA. Haplogroups are used to represent the major branch points on the mitochondrial phylogenetic tree. Understanding the evolutionary path of the female lineage has helped population geneticists trace the matrilineal inheritance of modern humans back to human origins in Africa and the subsequent spread across the globe.

However Balloux et al. (2009) have shown that mtDNA also correlates with climate and that temperature-based natural selection has helped shape global mtDNA patterns so that the assumption of pure genetic drift may be incorrect.

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 woman at the root of all these groups is the matrilineal most recent common ancestor (MRCA) for all currently living humans. She is commonly called Mitochondrial Eve.

Evolutionary relationship

Lineage Perspective

This phylogenetic tree is based on the Van Oven 2009 tree and subsequent published research.

- **L** (Mitochondrial Eve)
  - **L0**
  - L1-6
    - **L1**

- L2-6
  - **L5**
  - L2'3'4'6
    - **L2**
    - L3'4'6
      - **L6**
      - L3'4
        - **L4**
        - **L3**
          - **M**
            - M8: **CZ** (**C**, **Z**)
            - M9: **E**
            - M12'G: **G**
            - M29'Q: **Q**
            - **D**
          - **N**
            - N1: **I**
            - N2: **W**
            - N9: **Y**
            - **A**
            - **S**
            - **X**
            - **R**
              - R0 (FMKA pre-HV)
                - **HV**: (**H**, **V**)
              - pre-JT or R2'JT
                - **JT**: (**J**, **T**)
              - R9: **F**
              - R11'B: **B**
              - **P**
              - **U** (formerly UK)
                - **U8**: **K**

Chronological development of haplogroups

European haplogroups

Bryan Sykes had claimed there were seven major mitochondrial lineages for modern Europeans but others now put the number at 10-12. These additional "daughters"

generally include haplogroups I, M and W. A recent paper re-mapped European haplogroups as H, J, K, N1, T, U4, U5, V, X and W.

- N : 75,000 years ago (in North-East Africa)
- R : 70,000 years ago (in South-West Asia)
- U : 60,000 years ago (in North-East Africa or South-West Asia)
- pre-JT : 55,000 years ago (in the Middle East)
- JT : 50,000 years ago (in the Middle East)
- U5 : 50,000 years ago (in Western Asia)
- U6 : 50,000 years ago (in North Africa)
- U8 : 50,000 years ago (in Western Asia)
- pre-HV : 50,000 years ago (in the Near East)
- J : 45,000 years ago (in the Near East or Caucasus)
- HV : 40,000 years ago (in the Near East)
- H : over 35,000 years ago (in the Near East or Southern Europe)
- X : over 30,000 years ago (in north-east Europe)
- U5a1 : 30,000 years ago (in Europe)
- I : 30,000 years ago (Caucasus or north-east Europe)
- J1a : 27,000 years ago (in the Near East)
- W : 25,000 years ago (in north-east Europe or north-west Asia)
- U4 : 25,000 years ago (in Central Asia)
- J1b : 23,000 years ago (in the Near East)
- T : 17,000 years ago (in Mesopotamia)
- K : 16,000 years ago (in the Near East)
- V : 15,000 years ago (arose in Iberia and moved to Scandinavia)
- H1b : 13,000 years ago (in Europe)
- K1 : 12,000 years ago (in the Near East)
- H3 : 10,000 years ago (in Western Europe)

Chapter- 6

Human Mitochondrial Molecular Clock

The **human mitochondrial molecular clock** is the rate at which mutations have been accumulating in the mitochondrial genome of hominids during the course of human evolution. The archeological record of human activity from early periods in human prehistory is relatively limited and its interpretation has been controversial. Because of the uncertainties from archeological record, scientists have turned to molecular dating techniques in order to refine the timeline of human evolution. A major goal of scientists in the field is to develop an accurate hominid mitochondrial molecular clock which could then be used to confidently date events that occurred during the course of human evolution.

Estimates of the mutation rate of human mitochondrial DNA (mtDNA) vary greatly depending on the available data and the method used for estimation. The two main methods of estimation, phylogeny based methods and pedigree based methods, have produced mutation rates that differ by almost an order of magnitude. Current research has been focused on resolving the high variability obtained from different rate estimates.

Rate variability

A major assumption of the molecular clock theory is that mutations within a particular genetic system occur at a statistically uniform rate and this uniform rate can be used for dating genetic events. In practice the assumption of a single uniform rate is an oversimplification. Though a single mutation rate is often applied, it is often a composite or an average of several different mutation rates. Many factors influence observed mutation rates and these factors include the type of samples, the region of the genome studied and the time period covered.

Actual vs. observed rates

The rate at which mutations occur during reproduction, the germline mutation rate, is thought to be higher than all observed mutation rates, because not all mutations are successfully passed down to subsequent generations. MtDNA is only passed down along the matrilineal line, and therefore mutations passed down to sons are lost. Random

genetic drift may also cause the loss of mutations. For these reasons, the actual mutation rate will not be equivalent to the mutation rate observed from a population sample.

Population size

Population dynamics are believed to influence observed mutation rates. When a population is expanding, more germline mutations are preserved in the population. As a result, observed mutation rates tend to increase in an expanding population. When populations contract, as in a population bottleneck, more germline mutations are lost. Population bottlenecks thus tend to slow down observed mutation rates. Since the emergence of the species homo sapiens about 200,000 years ago, human population have expanded from a few thousand individuals living in Africa to over 6 billion all over the world. However the expansion has not been uniform, the history of human populations may have consisted of both bottlenecks and expansions.

Structural variability

The mutation rate across the mitochondrial genome is not uniformly distributed. Certain regions of the genome are known to mutate more rapidly than others. The Hypervariable regions are known to be highly polymorphic relative to other parts of the genome.

The rate at which mutations accumulate in coding and non-coding regions of the genome also differs as mutations in the coding region are subject to purifying selection. For this reason, some studies avoid coding region or synonymous mutations when calibrating the molecular clock. Loogvali et al. (2009) only consider synonymous mutations, they have recalibrated the molecular clock of human mtDNA as 7990 years per synonymous mutation over the mitochondrial genome. Soares et al. (2009) consider both coding and non-coding region mutations to arrive at a single mutation rate, but apply a correction factor to account for selection in the coding region.

Temporal variability

The mutation rate has been observed to vary with time. Mutation rates within the human species are faster than those observed along the human-ape lineage. The mutation rate is also thought to be faster in recent times, since the beginning of the Holocene 11,000 years ago.

Parallel mutations and saturation

Parallel mutation (sometimes referred to as Homoplasy) or convergent evolution occurs when separate lineages have the same mutation independently occur at the same site in the genome. Saturation occurs when a single site experiences multiple mutations. Parallel mutations and saturation result in the underestimation of the mutation rate because they are likely to be overlooked.

Heteroplasmy

Indivdiduals affected by heteroplasmy have a mixture of mtDNA types, some with new mutations and some without. The new mutations may or may not be passed down to subsequent generations. Thus the presence of heteroplasmic individuals in a sample may complicate the calculation of mutation rates.

Methods

Pedigree based

Pedigree methods estimate the mutation rate by comparing the mtDNA sequences of a sample of parent/offspring pairs or analyzing mtDNA sequences of individuals from a deep-rooted genealogy. The number of new mutations in the sample is counted and divided by the total number of parent-to-child DNA transmission events to arrive at a mutation rate.

Phylogeny based

Phylogeny based methods are estimated by first reconstructing the haplotype of the most recent common ancestor (MRCA) of a sample of two or more genetic lineages. A requirement is that the time to the most recent common ancestor(TMRCA) of the sample of lineages must already be known from other independent sources, usually the archeological record. The average number of mutations that have accumulated since the MRCA is then computed and divided by the TMRCA to arrive at the mutation rate. The human mutation rate is usually estimated by comparing the sequences of modern humans and chimpanzees and then reconstructing the ancestral haplotype of the chimpanzee-human common ancestor. According to the paleontological record the last common ancestor of humans may have lived around 6 million years ago.

Pedigree vs. Phylogeny comparison

Rates obtained by pedigree methods are about 10 times faster than those obtained by phylogenetic methods. Several factors acting together may be responsible for this difference. As pedigree methods record mutations in living subjects, the mutation rates from pedigree studies are closer to the germline mutation rate. Pedigree studies use genealogies that are only a few generations deep whereas phylogeny based methods use timescales that are thousands or millions of years deep. According to Henn et al. 2009, phylogeny based methods take into account events that occur over long time scales and are thus less affected by stochastic fluctuations. Howell et al. 2003 suggests that selection, saturation, parallel mutations and genetic drift are responsible for the differences observed between pedigree based methods and phylogeny based methods.

Estimating based on AMH archaeology

Methods/parameters for archaeologically estimated dates of mitochondrial Eve			
Study	Sequence type	T_{Anchor} (location)	Referencing method (correction method)
Cann, Stoneking & Wilson (1987)	Restriction fragments	40, 30, and 12 Ka (Australia, New Guinea New World)	archaeologically defined migrations matched with estimated sequence divergence rates
Endicott & Ho (2008)	Genomic	40 to 55 Ka (Papua New Guinea) 14.5 to 21.5 Ka (Haps H1 and H3)	PNG following Haplogroup P

Anatomical modern humans (AMH) spread out of Africa and over a large area of Eurasia and left artifacts along the northern coast of the Southwest, South, Southeast and East Asia. Cann, Stoneking & Wilson (1987) did not rely on a predicted T_{CHLCA} to estimate SNP rates. Instead, they used evidence of colonization in Southeast Asia and Oceania to estimate mutation rates. In addition they used RFLP technology (*Restriction fragment length polymorphism*) to examine differences between DNA. Using these techniques this group came up with a T_{MRCA} of 140,000 to 290,000 years. It should be noted however that Cann et al. (1987) estimated the TMRCA of humans to be approximately 210 ky and the most recent estimates Soares et al. 2009 (using 7 million year chimpanzee human mtDNA MRCA) differ by only 9%, which is relatively close considering the wide confidence range for both estimates and calls for more ancient T_{CHLCA}.

Endicott & Ho (2008) have reevaluated the predicted migrations globally and compared those to the actual evidence. This group used the coding regions of sequences. They postulate that the molecular clock based on chimp-human comparisons is not reliable, particularly in predicting recent migrations, such as founding migrations into Europe, Australia, and the Americans. With this technique this group came up with a T_{MRCA} of 82,000 to 134,000 years.

The anchoring method used by Cann et al. (1987), is based on contemporary understanding of archaeology, which places humans in East Asia by 40 kya. However it is currently known that anatomically modern humans reached Southwestern China well before 42,000 years ago (ka). In addition, the dates of the Mungo Lake remains have been reestimated to between 42 and 63 Ka consistent with other recent evidence for earlier occupation. There is evidence of human occupation in India from 76 Ka, and the arguably anatomically modern human remains at Jebel Qafzeh have been reestimated to 93 ka. The underestimate of D-loop divergence rate resulted in an overestimate of the TMRCA while the underestimate of the age of human migration from Africa resulted in an underestimate, such that the errors largely balanced each other.

Estimating based on CHLCA

Because chimps and humans share a matrilineal ancestor, establishing the geological age of that last ancestor allows the estimation of the mutation rate. The chimp-human last common ancestor (CHLCA) is frequently applied as an anchor for mt-T_{MRCA} studies with ranges between 4 and 13 million years cited in the literature. This is one source of variation in the time estimates. The other weakness is the non-clocklike accumulation of SNPs, would tend to make more recent branches look older than they actually are.

SNP rates as described by Soares et al. (2009)		
Regions(s)	Subregions (or site within codon)	SNP rate (per site * year)
Control region	HVR I	1.6×10^{-7}
	HVR II	2.3×10^{-7}
	remaining	1.5×10^{-8}
Protein-coding	(1st and 2nd)	8.8×10^{-9}
	(3rd)	1.9×10^{-8}
DNA encoding rRNA (rDNA)		8.2×10^{-9}
DNA encoding tRNA (tDNA)		6.9×10^{-9}
other		2.4×10^{-8}
T_{CHLCA} assumed 6.5 Ma, relative rate to 1st & 2nd codons		

These two sources may balance each other or amplify each other depending on the direction of the T_{CHLCA} error. There are two major reasons why this method is widely employed. First the pedigree based rates are inappropriate for estimates for very long periods of time. Second, while the archaeology anchored rates represent the intermediate range, archaeological evidence for human colonization often occurs well after colonization. For example, colonization of Eurasia from west to east is believed to have occurred along the Indian Ocean. However, the oldest archaeological sites that also demonstrate anatomically modern humans (AMH) are in China and Australia, greater than 42,000 years in age. However the oldest Indian site with AMH remains is from 34,000 years, and another site with AMH compatible archaeology is in excess of 76,000 years in age. Therefore application of the anchor is a subjective interpretation of when humans were first present.

A simple measure the sequence divergence between humans and chimps by observing the SNPs. Given that the mitogenome is about 16553 base pairs in length (each base-pair which can be aligned with known references is called a site). The formula is:

$$rate = \frac{SNPs}{(2T_{CHLCA}16553)}$$

The '2' in the denominator is derived from the 2 lineages, human and chimpanzee, that split from the CHLCA. Ideally it represents the accumulation of mutations on both

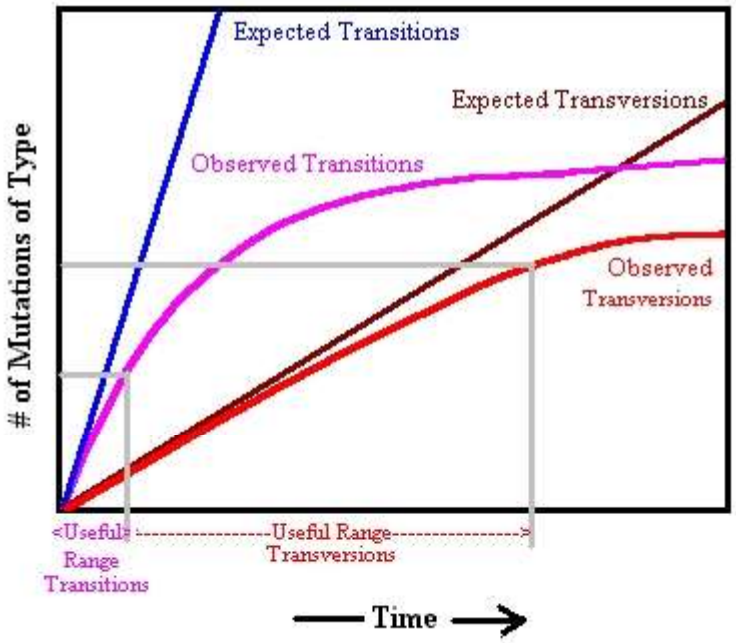
lineages but in different positions (SNPs). As long as the number of SNP observed approximates the number of mutations this formula works well. However, at rapidly evolving sites mutations are obscured by saturation affects. Sorting positions within the mitogenome by rate and compensating for saturation are alternative approaches.

Because the T_{CHLCA} is subject to change with more paleontological information, the equation described above allows the comparison of TMRCA from different studies.

Methods/parameters for estimating date of mitochondrial Eve			
Study	Sequence type	T_{CHLCA} (sorting time)	Referencing method (correction method)
Vigilant et al. (1991)	HVR	4 to 6 Ma	CH transversions, (15:1 transition:transversion)
Ingman et al. (2000)	genomic (not HVR)	5 Ma	CH genomic comparison
Endicott & Ho (2008)	genomic (not HVR)	5 to 7.5 Ma	CH (relaxed rate, rate-class defined)
Gonder et al. (2007)	genomic (not HVR)	6.0 Ma (+ 0.5 Ma)	CH (rate class defined)
Mishmar et al. (2003)	genomic (not HVR)	6.5 Ma (+ 0.5 Ma)	CH (rate class defined)
Soares et al. (2009)	genomic	6.5Ma (+ 0.5 Ma)	CHLCA anchored, (Examined selection by Ka/(Ks + k))
Chimpanzee to Human = CH, LCA = last common ancestor			

Early, HVR, sequence based methods

To overcome the affects of saturation, HVR analysis relied on the transversional distance between humans and chimpanzees. A transition to transversion ratio was applied to this distance to estimate sequence divergence in the HVR between chimpanzees and humans, and divided by an assumed T_{CHLCA} of 4 to 6 million years. Based on 26.4 substitutions between chimpanzee and human and 15:1 ratio, the estimated 396 transitions over 610 base-pairs demonstrated sequence divergence of 69.2% (rate * T_{CHLCA} of 0.369), producing divergence rates of *roughly 11.5% to 17.3% per million years*.



HVR is exceptionally prone to saturation, leading to the underestimation of the SNP rate when comparing very distantly related lineages

Vigilant et al. (1991) also estimated the sequence divergence rate for the sites in the rapidly evolving HVR I and HVR II regions. As noted in the table above, the rate of evolution is so high that site saturation occurs in direct chimpanzee and human comparisons. Consequently this study used transversions, which evolve at a slower rate than the more common transition polymorphisms. Comparing chimp and human mitogenomes, they noted 26.4 transversions within the HVR regions, however they made no correction for saturation. As more HVR sequence was obtained following this study, it was noted that the dinucleotide site CRS:16181-16182 experienced numerous transversions in parsimony analysis, many of these were considered to be sequencing errors. However the sequencing of Feldhofer I Neanderthal revealed that there was also a transversion between humans and Neanderthals at this site. In addition, Soares et al. (2009) noted three sites in which recurrent transversions had occurred in human lineages, two of which are in HVR I, 16265 (12 occurrences) and 16318(8 occurrences). Therefore, 26.4 transversions was an underestimate of the likely number of transversion events. The year 1991 study also used a transition-to-transversion ratio from the study of old world monkeys of 15:1. However, examination of chimp and gorilla HVR reveals a rate that is lower, and the examination of humans places the rate at 34:1. Therefore this study underestimated that level of sequence divergence between chimpanzee and human. The estimated sequence divergence 0.738/site (includes transversions) is significantly lower than the ~2.5 per site suggested by Soares et al. (2009). These two errors would result in an overestimate of the human mitochondrial TMRCA. However, they failed to detect the basal L0 lineage in the analysis and also failed to detect recurrent transitions in many lineages, which also underestimate the TMRCA. Also, Vigilant et al. (1991) used a more recent CHLCA anchor of 4 to 6 million years.

Coding region sequence based methods

Partial coding region sequence originally supplemented HVR studies because complete coding region sequence was uncommon. There were suspicions that the HVR studies had missed major branches based on some earlier RFLP and coding region studies. Ingman et al. (2000) was the first study to compare genomic sequences for coalescence analysis. Coding region sequence discriminated M and N haplogroups and L0 and L1 macrohaplogroups. Because the genomic DNA sequencing resolved the two deepest branches it improved some aspects estimating TMRCA over HVR sequence alone. Excluding the D-loop and using a 5-million-year T_{CHLCA}, Ingman et al. (2000) estimated the mutation rate *to be* 1.70×10^{-8} *per site per year* (rate * T_{CHLCA} = 0.085, 15,435 sites).

However, coding region DNA has come under question because coding sequences are either under purifying selection to maintain structure and function, or under regional selection to evolve new capacities. The problem with mutations in the coding region has been described as such: mutations occurring in the coding region that are not lethal to the mitochondria can persist but are negatively selective to the host; over a few generations these will persist, but over thousands of generations these slowly are pruned from the population, leaving SNPs. However, over thousands of generations regionally selective mutations may not be discriminated from these transient coding region mutations. The problem with rare mutations in the human mitogenomes is significant enough to prompt a half-dozen recent studies on the matter.

Ingman et al. (2000) estimated the *non-D loop region* evolution 1.7×10^{-8} per year per site based on 53 non-identical genomic sequence overrepresenting Africa in a global sample. Despite this over-representation, the resolution of the L0 subbranches was lacking and one other deep L1 branches has been found. Despite these limitations that sampling was adequate for the hallmark study. Today, L0 is restricted to African populations, whereas L1 is the ancestral haplogroup of all non-Africans, as well as most Africans. Mitochondrial Eve's sequence can be approximated by comparing a sequence from L0 with a sequence from L1. By reconciling the mutations in L0 and L1. The mtDNA sequences of contemporary human populations will generally differ from Mitochondrial Eve's sequence by about 50 mutations. Mutation rates were not classified according to site (other than excluding the HVR reigons). The T_{CHLCA} used in the year 2000 study of 5 Ma was also lower than values used in the most recent studies.

Inter-comparing rates and studies

Molecular clocking of mitochondrial DNA has been criticized because of its inconsistent molecular clock. A retrospective analysis of any pioneering process will reveal inadequacies. With mitochondrial the inadequacies are the argument from ignorance of rate variation and overconfidence concerning the T_{CHLCA} of 5 Ma. Lack of historical perspective might explain the second issue, the problem of rate variation is something that could only be resolved by the massive study of mitochondria that followed. The number of HVR sequences that have accumulated from 1987 to 2000 increased by magnitudes. Soares et al. (2009) used 2196 mitogenomic sequences and uncovered

10,683 substitution events within these sequences. Eleven of 16560 sites in the mitogenome produced greater than 11% of all the substitutions with statistically significant rate variation within the 11 sites. They argue that there is a neutral-site mutation rate which is a magnitude slower than rate observed for the fastest site, CRS 16519. Consequently, purifying selection aside, the rate of mutation itself varies between sites, with a few sites much more likely to undergo new mutations relative to others. Soares et al. (2009) noted two spans of DNA, CRS 2651-2700 and 3028-3082, that had no SNPs within the 2196 mitogenomic sequences.

Chapter- 7

Human Modal Haplotypes

Haplogroup I1 (Y-DNA)

Haplogroup I1

Possible time of origin	4,000 to 20,000 BC
Possible place of origin	Scandinavia

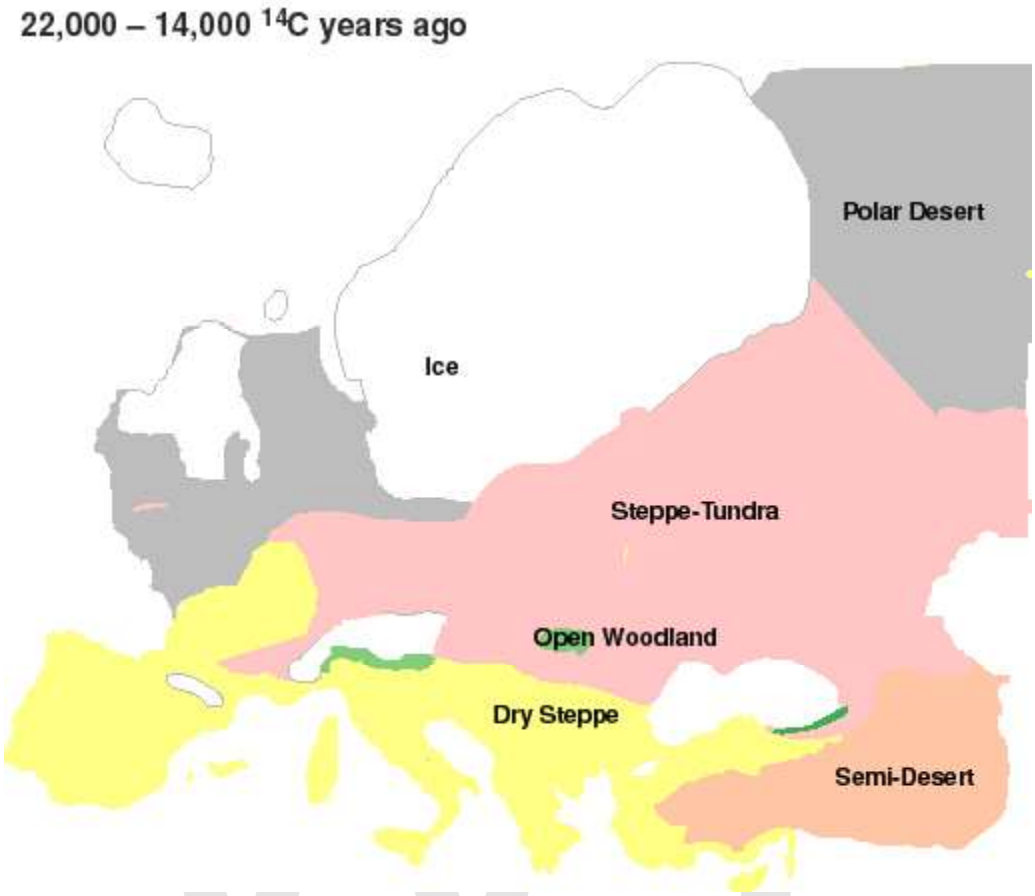
Ancestor	I
Defining mutations	M253, M307, P30, P40
Highest frequencies	People of Northern Europe (Norwegian, Swedish, Danish, Finnish, Sami, Estonian, German, Dutch, English, Scottish, Irish), French

In human genetics, **Haplogroup I1** is a Y chromosome haplogroup occurring at greatest frequency in Scandinavia, associated with the mutations identified as M253, M307, P30, and P40. These are known as single nucleotide polymorphisms (SNPs). It is a subclade of Haplogroup I. Before a reclassification in 2008, the group was known as **Haplogroup I1a**. Some individuals and organizations continue to use the I1a designation.

The group displays a very clear frequency gradient, with a peak of approximately 40 percent among the populations of western Finland and more than 50 percent in the province of Satakunta, around 35 percent in southern Norway, southwestern Sweden especially on the island of Gotland, and Denmark, with rapidly decreasing frequencies toward the edges of the historically Germanic sphere of influence.

Origins

For several years the prevailing theory was that during the Last Glacial Maximum (LGM) the predecessors of the I1 group sought refuge in the Balkans. For a time, the Ukraine was considered as an alternative. Yet, The Genographic Project claims that the founder of the I1 branch lived on the Iberian Peninsula during the LGM. Some have given southern France and the Italian peninsula as possible sites as well. Although the locations vary, proponents of the refuge theories do seem to agree on one issue: that the I1 subclade is from 15,000 to 20,000 years old.



Approximately 20,000 years ago, much of Europe was covered in ice and permafrost. People in Europe were forced south by the changing climate and topography. The European LGM refuges included the Iberian peninsula and the Balkans, where some believe the ancestors of I1 lived. The theory has been challenged recently by an opposing argument that I1 was not in existence during the LGM. Two primary cultures have been identified during this time: the Solutrean (Iberia and southern France) and the Gravettian (Balkans, Italy and Ukraine).

However, professor Ken Nordtvedt of Montana State University believes that I1 is a more recent group, probably emerging *after* the LGM. Other researchers including Peter A. Underhill of the Human Population Genetics Laboratory at Stanford University have since confirmed this hypothesis in independent research.

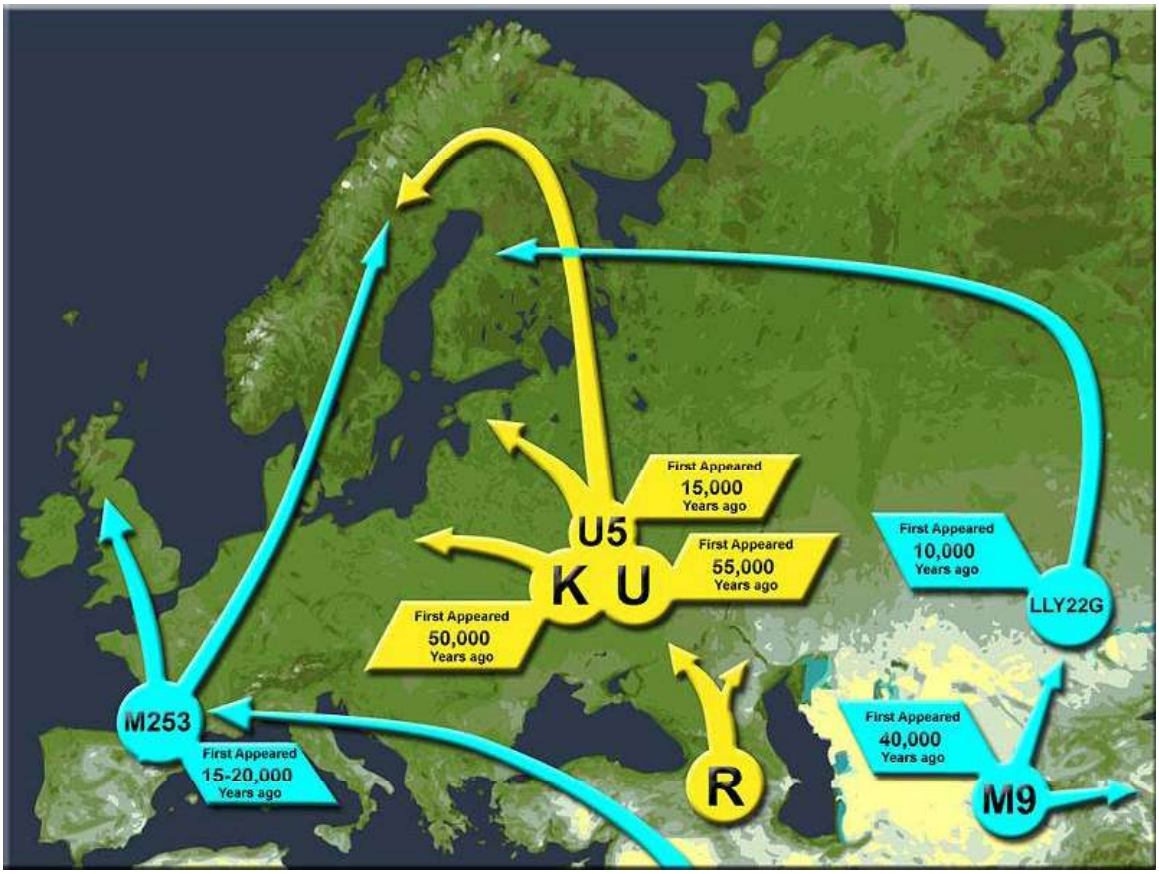
The study of I1, which some had argued was largely ignored by the genetic testing industry in favor of "mega-haplogroups" like R, is in flux. Revisions and updates to previous thinking, primarily published in academic journals, is constant, yet slow, showing an evolution in thought and scientific evidence.

The most recent common ancestor (MRCA) of I1 lived from 4,000 to 6,000 years ago somewhere in the far northern part of Europe, perhaps Denmark, according to Nordtvedt. His descendants are primarily found among the Germanic populations of northern Europe

and the bordering Uralic and Celtic populations, although even in traditionally German demographics I1 is overshadowed by the more prevalent Haplogroup R.

When SNPs are unknown or untested and when short tandem repeat (STR) results show eight allele repeats at DNA Y chromosome Segment (DYS) 455, haplogroup I1 can be predicted correctly with a very high rate of accuracy, 99.3 to 99.8 percent, according to Whit Athey and Vince Vizachero. This is almost exclusive to and ubiquitous in the I1 haplogroup, with very few having seven, nine, or another number. Furthermore, DYS 462 divides I1 geographically. Nordtvedt considers 12 allele repeats to be more likely Anglo-Saxon and on the southern fringes of the I1 map, while 13 signifies more northerly, Nordic origins. Nordtvedt has repeatedly argued that, at least for I1, SNP testing is generally not as beneficial as expanded STR results.

Subclades



One theory showing the dispersal of I1 (M253) in Europe based on information from The Genographic Project.

Note: The systematic subclade names have changed several times in recent years, and are likely to change again, as new markers which clarify the sequence of branchings of the tree are discovered.

- I1 (M253, M307, L75, L80, L81, L118, L121, L123, L125, M450, M307.1/P203.1, P30, P40, S62, S63, S64, S65, S66, S107, S108, S109, S110, S111) *formerly I1a*
    - I1*
    - I1a (M21) *formerly I1a2*
    - I1b (M227) *formerly I1a1, I1a4*
        - I1b*
        - I1b1 (M72) *formerly I1a1a, I1a3*
    - I1c (P259) *formerly I1d*
    - I1d (L22/S142)
        - I1d*
        - I1d1 (P109) *formerly I1c*
    - I1e (S79)
    - L338

Distribution

Outside Scandinavia, distribution of Haplogroup I1 is closely correlated with Haplogroup I2a1, but among Scandinavians including both Germanic and Uralic peoples of the region nearly all of the Haplogroup I chromosomes are I1. It is common near the southern Baltic and North Sea coasts, although successively decreasing the further south geographically. The Migration Period or "wandering of peoples" may explain the dispersion of I1 into areas beyond northern Europe.

Britain

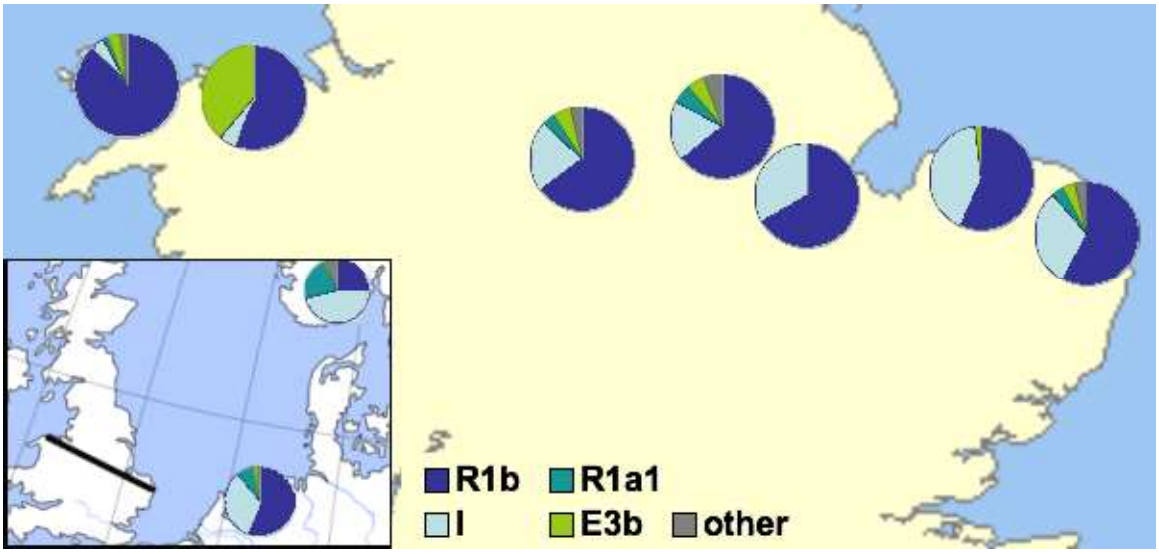
A map of England showing the general locations of the Germanic (Angle, Saxon and Jute) and indigenous (Briton, Hwicca) peoples around the year 600

The traditional view of British and Irish prehistory was that several waves of migration had resulted in widespread, if not total, population displacement. After the Last Glacial Maximum the region was first repopulated by Paleolithic hunter gatherers. During the Neolithic period, with the spread of farming, this population was supposedly replaced by the farmers. Later immigrations were thought to have accompanied the transitions to bronze and iron-working, known respectively as the Bronze and Iron Ages. The introduction of iron was particularly significant because archaeologists had associated it with the Hallstatt and La Tène cultures. These came to be associated by early archaeologists with the so-called Celtic culture, which was seemingly widespread on the continent. A later migration, that of the Anglo-Saxons, was also claimed to have led to total population replacement, but genetic evidence suggests otherwise. In his

Ecclesiastical History of the English People, Bede claimed that the Angles came to Britain en masse as an entire nation leaving no one behind in their homeland Angeln. Other population movements (though not total displacements) recorded during historical times include those of the Danish and Norwegian Vikings, Danes in the east of England (especially the Danelaw) and Norwegians in the Shetland and Orkney Isles, Western Isles and Ireland.

The replacement model has been under sustained attack since the 1960s, with researchers asserting a much greater continuity than previously known or acknowledged. British archaeologist Simon James attributes the idea of large-scale mass migration to the assumption of primitivism about earlier inhabitants, assuming that cultural changes, such as nomadic hunter-gathering to farming, stone-working to metalworking, and bronze-working to iron-working, required newcomers introducing materials and techniques to the indigenous population, rather than them learning through trade or other methods.

Francis Pryor has stated that he "can't see any evidence for *bona fide* mass migrations after the Neolithic." Historian Malcolm Todd writes, "It is much more likely that a large proportion of the British population remained in place and was progressively dominated by a Germanic aristocracy, in some cases marrying into it and leaving Celtic names in the, admittedly very dubious, early lists of Anglo-Saxon dynasties. But how we identify the surviving Britons in areas of predominantly Anglo-Saxon settlement, either archaeologically or linguistically, is still one of the deepest problems of early English history." Although the idea of mass human migrations into Great Britain and Ireland is now a relatively minor point of view amongst British and Irish archaeologists, there is still a perception outside of the archaeological community that an "Anglo-Saxon" mass migration (especially) occurred, and that this forms a fundamental division between English "Anglo-Saxon" populations in Great Britain and non-English "Celtic" populations.



Map showing the distribution of Y chromosomes in a trans section of England and Wales from the paper "Y Chromosome Evidence for Anglo-Saxon Mass Migration". The

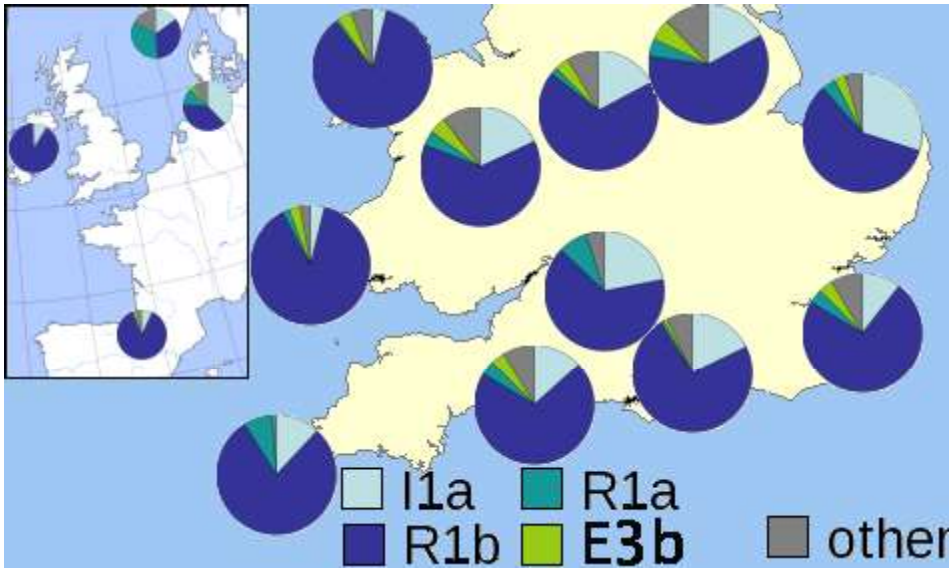
authors attribute the differences in frequencies of haplogroup I to Anglo-Saxon mass migration into England, but not into Wales.

In 2002 a paper titled "Y Chromosome Evidence for Anglo-Saxon Mass Migration" was published by the Centre for Genetic Anthropology at the University College London in cooperation with Vrije Universiteit and the University of California, Davis claiming direct genetic evidence for population differences between the English and Welsh populations and proposed a model for mass invasion of eastern Great Britain from northern Germany and Denmark. The authors assumed that populations with large proportions of haplogroup I originated from northern Germany or southern Scandinavia, particularly Denmark, and that their ancestors had migrated across the North Sea with Anglo-Saxon migrations and Danish Vikings.

In her book *Origins of the English* Catherine Hills criticized these conclusions, arguing that a biased sampling strategy flawed the study, especially since testing was limited only to regions in England where Danes were known to have settled during the Danelaw, which is archaeologically distinct. In the paper the main claim by the researchers was

that an Anglo-Saxon immigration event affecting 50–100% of the Central English male gene pool at that time is required. We note, however, that our data do not allow us to distinguish an event that simply added to the indigenous Central English male gene pool from one where indigenous males were displaced elsewhere or one where indigenous males were reduced in number … This study shows that the Welsh border was more of a genetic barrier to Anglo-Saxon Y chromosome gene flow than the North Sea … These results indicate that a political boundary can be more important than a geophysical one in population genetic structuring.

The paper was widely publicized in the media, especially in the United Kingdom, but reporting was often misleading and inaccurate. For example, the BBC claimed that the "English and Welsh are races apart" and asserted "that between 50% and 100% of the indigenous population of what was to become England was wiped out" though this was not a claim of the paper. The conclusion for evidence of mass Anglo-Saxon migration, and that east English samples were more similar to Frisian samples than to Welsh samples, did not support the archaeological orthodoxy of modern times. A year later, in 2003, the paper "A Y Chromosome Census of the British Isles" was published by Capelli *et al.*. This paper, which sampled Great Britain and Ireland on a grid, found a much smaller difference between Welsh and English samples, and was much more characterised by isolation by distance, with a gradual decrease in Haplogroup I frequency moving westwards in southern Great Britain. It also found North German and Danish samples were not more similar to east English samples than Welsh samples.



Distribution of Y chromosome haplogroups from Capelli *et al.* (2003). Haplogroup I is present in all populations, with higher frequencies in the east and lower frequencies in the west. There appears to be no discrete boundary as observed by Weale *et al.* (2002)

Oxford archaeologist David Miles has argued that 80 percent of the genetic makeup of native Britons probably comes from "just a few thousand" nomadic tribesmen who arrived 12,000 years ago, at the end of the Ice Age. This suggests later waves of immigration may have been too small to have significantly affected the genetics of the pre-existing population.

Traditionally, areas with a majority Angle influence included the Kingdoms of Northumbria (*Nord Angelnen, Nordimbria Anglorum*), East Anglia (*Ost Angelnen*) and Mercia (*Mittlere Angelnen*) while the Saxon areas were the Kingdoms of Sussex (*Suth Seaxe*), Essex (*Est Seaxna*), and Wessex (*West Seaxna*). The Kingdom of Kent was considered a place of another Germanic tribe, the Jutes. Stephen Oppenheimer suggested that the Anglo-Saxon invasions actually had been predominantly Anglian.

Meanwhile, Bryan Sykes has said that the Anglo-Saxons made a substantial contribution to the genetic makeup of England, but probably less than 20 percent of the total, even in southern England, where raids and settlements were supposedly commonplace. His conclusions, on Britain at least, mirror those of other researchers including Siiri Rootsi and Nordtvedt. A report on the Saxons who were part of the Germanic settlement of Britain during and after the fifth century was issued by University College London in July 2006, with a wide-ranging estimate for the total number of settlers varying between 10,000 and 200,000.

The Vikings, both Danes and Norwegians, also made a substantial contribution after the Angles, Saxons and Jutes, Sykes said, with concentrations in central, northern, and eastern England, territories of the ancient Danelaw. Sykes said he found evidence of a very heavy Viking contribution in the Orkney and Shetland Islands, near 40 percent.

Mitochondrial DNA as well as Y DNA of northern Germanic origin was discovered at substantial rates in all of these areas, showing that the Vikings engaged in large-scale settlement, Sykes explained. However, Nordtvedt has said that separating I1 haplotypes into Viking and non-Viking groups has been impossible thus far.

Evidence of Norman genetic influence in England was extremely small – about two percent according to Sykes, discounting the idea that William the Conqueror, his troops and any settlers disrupted and displaced previous cultures. Some notable British historians and Anglophiles including J. R. R. Tolkien assumed that the Norman invasion of AD 1066 greatly affected the society of the time and that little survived from the "original" Britons. This worldview permeates Tolkien's *The Lord of the Rings* and other writings, though he focuses on Germanic folktales and legends rather than the Celtic in creating a replacement mythology, albeit fictional. In England, from the fifth to seventh centuries, the Anglo-Saxons soon developed their own variety as well.

Map of England from the *Anglo-Saxon Chronicle*

The study of languages and place names provides more supporting evidence. For example, Old English emerged from the Anglo-Frisian dialects brought to Britain by Germanic settlers and perhaps Roman soldiers. The convergence of varying languages lends credence to a diverse genetic pool. Initially, the English language began as a diverse group of dialects reflecting the varied backgrounds of the Anglo-Saxon kingdoms. One of these dialects, Late West Saxon, eventually dominated.

Then two waves of invaders brought new influences. The first was by language speakers of the Scandinavian branch, known as North Germanic. They conquered and colonized parts of Britain in the eighth and ninth centuries. The second was the Normans in the eleventh century, who spoke Old French and ultimately developed an English variety called Anglo-Norman. These two invasions caused English to become linguistically "mixed" to some degree. English developed into a "borrowing" language of great flexibility with a large vocabulary.

In England the Viking Age began dramatically on June 8, 793, when Norsemen destroyed the abbey at Lindisfarne, plundering and murdering indiscriminately. An incident four years earlier, in which three Viking ships were beached in Portland Bay, perhaps on a trading expedition, created some tension, but Lindisfarne was different. The devastation of Northumbria's Holy Island shocked many, including the royal Courts of Europe. More than any other single event, the attack on Lindisfarne cast a shadow on the perception of the Vikings for the next twelve centuries.

France

Genetic remnants remain in northern France, indicating a small influx of I1 men, likely during Viking raids and subsequent settlement. Subtle increases in I1 haplotypes indicate a modest contribution, perhaps from a combination of the Frankish migration during the last days of the Roman Empire and later Viking incursions. Nordtvedt subscribes to this concept.

The Franks, for whom France (literally "Land of the Franks") is named, were a Germanic tribe first identified in the third century as an ethnic group living north and east of the Lower Rhine. They founded one of the Germanic monarchies which replaced the Western Roman Empire from the fifth century. The Frankish state consolidated its hold over large parts of Western Europe by the end of the eighth century. The Carolingian Empire and its successor states were Frankish. French nobility were often descended from Frankish and Norman Germanic lineages and often bore Germanic names such as Charles de Gaulle, though his family Y chromosome DNA has not been tested. The name “de Gaulle” likely came from “De Walle” which in German means “the wall" of a fortification or city.

Following the successful example of a Cornish-Viking alliance in 722 at the Battle of Hehil, which helped stop the Anglo-Saxon conquest of Cornwall at the time, the people of Brittany (Bretons) made friendly overtures to the Danish Vikings in an effort to counter Frankish expansionism. In 866 the Vikings and Bretons united to defeat a

Frankish army at the Battle of Brissarthe, resulting in formal recognition of Brittany's independence.

The Vikings continued to tactically help their Breton allies by devastating Frankish areas under the Carolingians with pillaging raids. In 885, one of the minor Viking leaders named Rollo helped in the siege of Paris under the command of Danish king Sigfred. When Sigfred retreated in return for tribute the following year, Rollo stayed behind and was eventually bought off and sent to bother Burgundy by the Frankish king, Charles the Simple. Later, he returned to the Seine with a group of Danish followers who were called "Men of the North" or Norsemen. They invaded the area of northern France now known as Normandy.

Rather than pay Rollo to leave, as was customary, Charles the Simple realized that his armies could not effectively defend against the raids and guerrilla tactics, and decided to appease Rollo by giving him land and hereditary titles under the condition that he defend against other Vikings. Led by Rollo, the Vikings settled in Normandy after being granted the land. They subsequently established the Duchy of Normandy. The descendants who emerged from the interactions of Vikings, Franks, and Gallo-Romans became known as Normans. This may explain why a noticeably higher than average rate of men living in northwestern France today are I1.

The Scandinavian colonisation of Normandy was principally Danish, complemented by a strong Norwegian contingent, although a few Swedes were present. The Viking colonization was not a mass phenomenon, but they established themselves rather densely in some areas, particularly Pays de Caux and the northern part of the Cotentin. Toponymic and linguistic evidence supports this theory. The merging of the Scandinavian and native populations contributed to the creation of one of the most powerful feudal states of Western Europe. The naval ability of the Normans would allow them to conquer England, and participate in the Crusades.

Russia

From the ninth and into the tenth centuries, Scandinavian raiders and merchants travelled to Russia, Belarus and Ukraine, known as Varangians by the Byzantines. The Varangians have been described as a warrior elite or nobility.

Varangian leader Rurik is credited with founding the first Rus state. Although recent genetic studies have identified two major royal lines within Russian society, R1a and N1c1a, genetic research shows significant I1 contribution centering on Moscow.(dead link)

John Haywood, author of *The Great Migrations*, believes that a group known as the Rus preceded the Varangians. However, most identify the Rus people as a particular Varangian tribe. A large burial mound in Novgorod Oblast, Russia, known as a tumulus and dating from the ninth century, is similar to those found in Old Uppsala, Sweden. It is

reportedly well-defended against potential looters and has never been excavated. Local residents refer to it as 'Rurik's Grave'.

Scandinavians remained in control of areas such as Kiev until at least the mid-eleventh century. They became the nucleus of the Rus state, whose Golden Age in the eleventh and early twelfth centuries came to an abrupt end with the Mongol invasion of 1240.

Their campaigns are commemorated on many runestones in both Norway and Sweden, among them the Greece Runestones and the Varangian Runestones. The last major expedition appears to have been the ill-fated expedition of Ingvar the Far-Travelled to Serkland, a region southeast of the Caspian Sea, commemorated by the Ingvar Runestones. What happened to the men is not known.

Greece and Turkey

Another branch of Varangians dominated the Byzantine Empire military elite for a time. This could be the precursor of spikes in I1 haplotypes in Turkey and Greece near Istanbul. A military unit known as the Varangian Guard was established by Emperor Basil II. After Rus military recruits helped him quell a rebellion, Basil II formed an alliance with Vladimir I of Kiev and organized the guard. New recruits from Sweden, Denmark, and Norway continued the Scandinavian predominance of the guard until the late eleventh century. So many Swedes left to enlist in the guard that a medieval Swedish law stated that no man could gain his inheritance while remaining in Greece. In *The History of the Crusades* author Steve Runciman noted that by the time of the Emperor Alexius, the Byzantine Varangian Guard was largely recruited from Anglo-Saxons in England and "others who had suffered at the hands" of the Vikings and the Normans.

Haplotypes

Modal

Ken Nordtvedt has given the following 'modal haplotypes' within the I1 haplogroup according to examples found in I1 populations. Many I1-Norse types have been found to be downstream of the P109 SNP, concretely defining it as a haplogroup subclade and giving further credence to Nordtvedt's method of haplotyping. Furthermore, SNP L22 has been discovered to be upstream of P109, encompassing all of P109s Norse types, additional Norse types without P109 as well as Ultra Norse types excluded from the pool of P109 positives.

Such haplotyping is necessary because currently more resolution of potential subclades through matching STR alleles exists than is available via testing for known subclade SNPs in haplogroup I1.

I1 Anglo-Saxon (I1-AS) *Has its peak gradient in the Germanic lowland countries: northern Germany, Denmark, the Netherlands, as well as England and old Norman regions of France.*

DYS number	385a	385b	388	389i	389ii	390	391	392	393	394	426	437	439	447	448	449	454	455	458	459a	459b
Haplotype	13	14	14	12	28	22	10	11	13	14	11	16	11	23	20	28	11	8	15	8	9

I1 Norse (I1-N) *Has its peak gradient in Sweden.*

DYS number	385a	385b	388	389i	389ii	390	391	392	393	394	426	437	439	447	448	449	454	455	458	459a	459b
Haplotype	14	14	14	12	28	23	10	11	13	14	11	16	11	23	20	28	11	8	15	8	9

I1 Ultra-Norse Type 1 (I1-uN1) *Has its peak gradient in Norway.*

DYS number	385a	385b	388	389i	389ii	390	391	392	393	394	426	437	439	447	448	449	454	455	458	459a	459b
Haplotype	14	15	14	12	28	23	10	11	13	14	11	16	11	23	20	28;29	11	8	15	8	9

Many other Nordtvedt haplotypes exist, and Nordtvedt has continually refined the haplogroup with more types as they become apparent as more I1 types are tested.

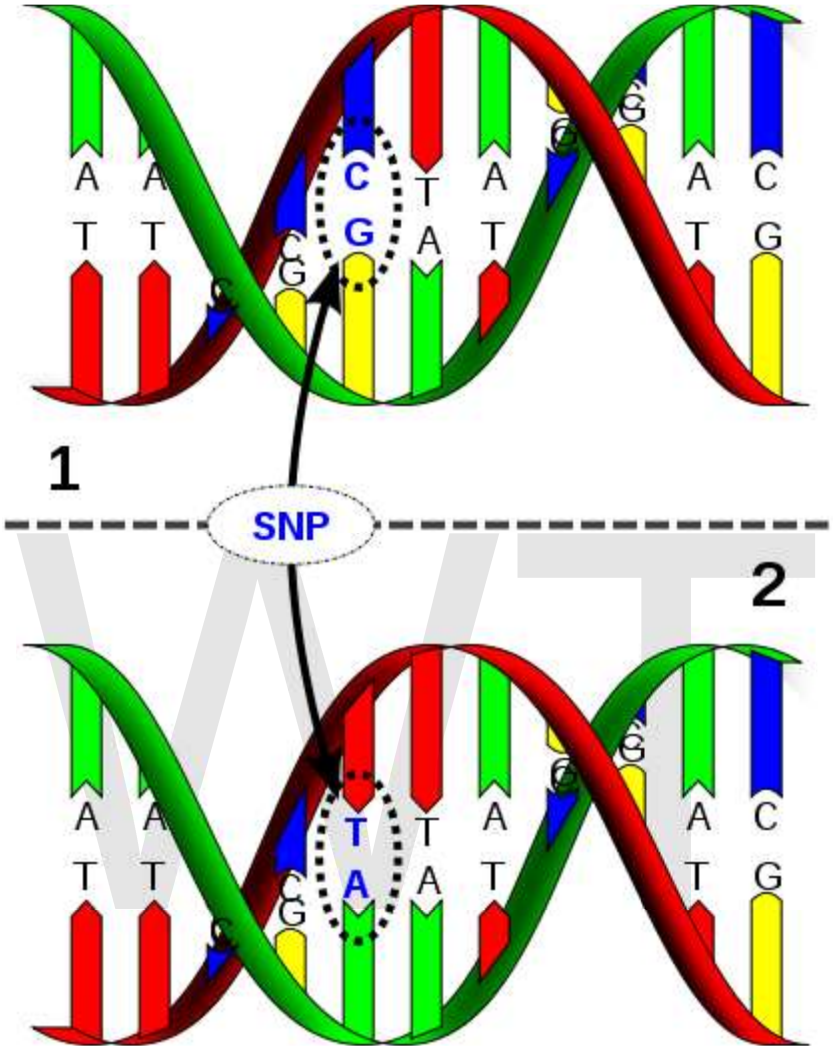
Famous

Alexander Hamilton, through genealogy and the testing of his descendants, has been placed within Y-DNA haplogroup I1.

DYS number	385a	385b	388	389i	389ii	390	391	392	393	394	426	437	439	447	448	449	454	455	458	459a	459b
Haplotype	13	14	14	13	29	22	10	11	13	14;15	11	16	12	22	20	31	11	8	15	8	9

DYS number	438	442	456	460	464a	464b	464c	464d	570	576	607	CDYa	CDYb	YCAIIa	YCAIIb
Haplotype	10	12	14	10	12	14	15	15	19	16	16	35	38	19	21

Mutations



DNA example: strand 1 differs from strand 2 at a single base pair location (a C >> T polymorphism).

The following are the technical specifications for known I1 haplogroup SNP and STR mutations.

Name: M253

Type: SNP
Source: M (Peter Underhill, Ph.D. of Stanford University)
Position: ChrY:13532101..13532101 (+ strand)
Position (base pair): 283
Total size (base pairs): 400
Length: 1

ISOGG HG: I1
Primer F (Forward 5′→ 3′): GCAACAATGAGGGTTTTTTTG
Primer R (Reverse 5′→ 3′): CAGCTCCACCTCTATGCAGTTT
YCC HG: I1
Nucleotide alleles change (mutation): C to T

Name: M307

Type: SNP
Source: M (Peter Underhill, Ph.D. of Stanford University)
Position: ChrY:21160339..21160339 (+ strand)
Length: 1
ISOGG HG: I1
Primer F: TTATTGGCATTTCAGGAAGTG
Primer R: GGGTGAGGCAGGAAAATAGC
YCC HG: I1
Nucleotide alleles change (mutation): G to A

Name: P30

Type: SNP
Source: PS (Michael Hammer, Ph.D. of the University of Arizona and James F. Wilson, D.Phil. at the University of Edinburgh)
Position: ChrY:13006761..13006761 (+ strand)
Length: 1
ISOGG HG: I1
Primer F: GGTGGGCTGTTTGAAAAAGA
Primer R: AGCCAAATACCAGTCGTCAC
YCC HG: I1
Nucleotide alleles change (mutation): G to A
Region: ARSDP

Name: P40

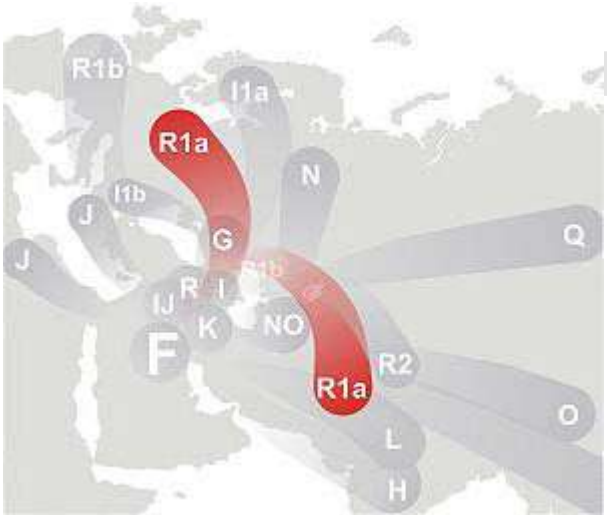
Type: SNP
Source: PS (Michael Hammer, Ph.D. of the University of Arizona and James F. Wilson, D.Phil. at the University of Edinburgh)
Position: ChrY:12994402..12994402 (+ strand)
Length: 1
ISOGG HG: I1
Primer F: GGAGAAAAGGTGAGAAACC
Primer R: GGACAAGGGGCAGATT
YCC HG: I1
Nucleotide alleles change (mutation): C to T
Region: ARSDP

Name: DYS455

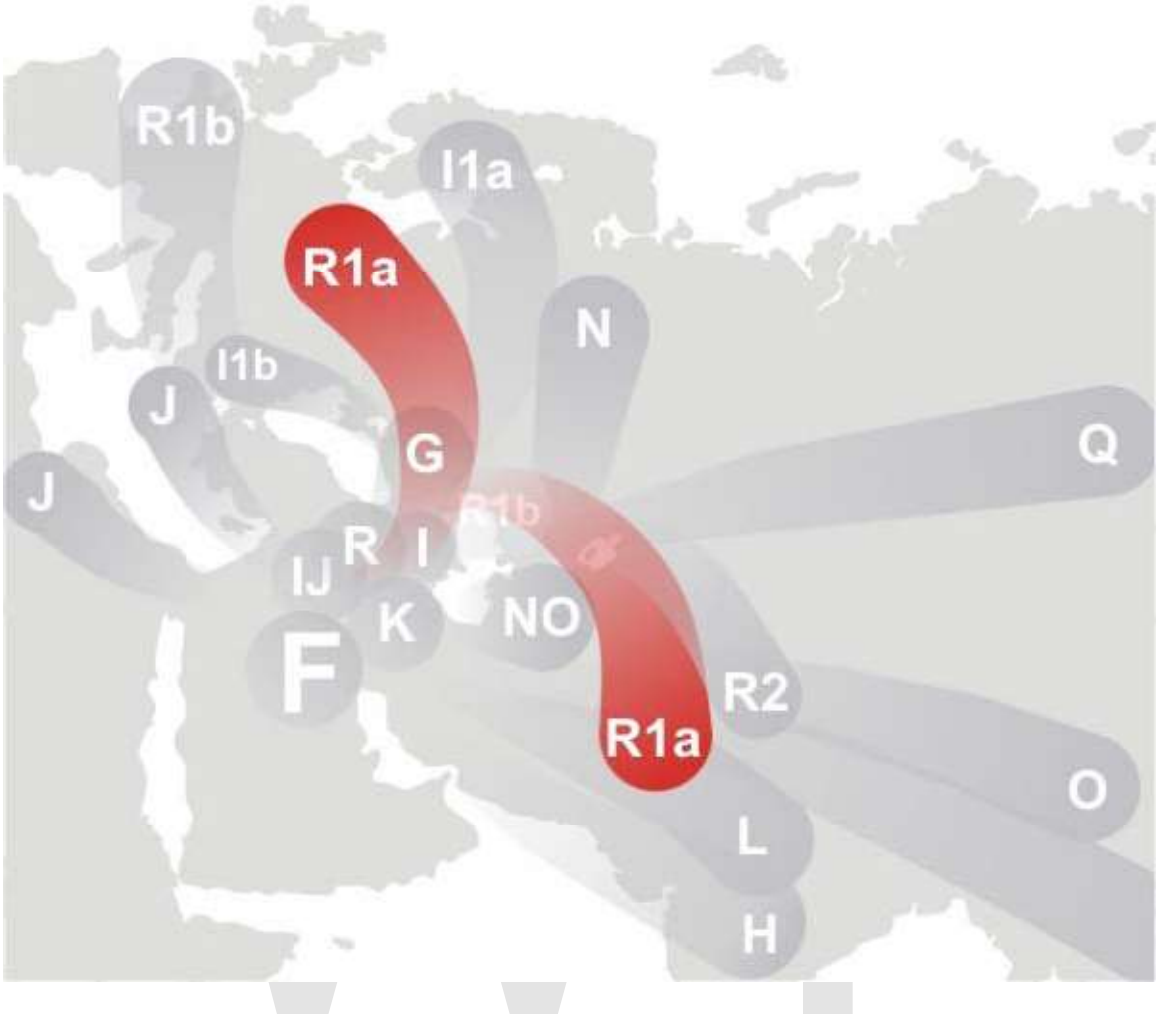
Type: STR (repeat)
Position: ChrY:6971459..6971638 (+ strand)
Length: 180
Primer F: ATCTGAGCCGAGAGAATGATA
Primer R: GGGGTGGAAACGAGTGTT

Haplogroup R1a (Y-DNA)

Haplogroup R1a



Possible time of origin	probably more recent than 18,500 years BP
Possible place of origin	Asia,most probably South Asia. Other possibilities include Central Asia, Middle East, and Eastern Europe.
Ancestor	R1 (R-M173)
Descendants	R1a1a1 to R1a1a8. R-M458 being the most significant (R1a1a7 in Underhill et al. (2009)).
Defining mutations	1. M420 now defines R1a in the broadest sense. 2. Within R1a, SRY1532.2 also known as SRY10831.2, now defines R1a1, previously R1a. 3. M17 and M198 (equivalent to one another) now define R1a1a, previously R1a1, and often referred to as if equal to R1a.
Highest frequencies	Parts of Eastern Europe, Scandinavia, Central Asia, Siberia and South Asia.



Haplogroup R1a is the phylogenetic name of a major clade of human Y-chromosome lineages. In other words, it is a way of grouping a significant part of all modern men according to a shared male-line ancestor. It is common in many parts of Eurasia and is frequently discussed in human population genetics and genetic genealogy. One sub-clade (branch) of R1a, currently designated R1a1a, is much more common than the others in all major geographical regions. R1a1a, defined by the SNP mutation M17, is particularly common in a large region extending from South Asia and Southern Siberia to Central Europe and Scandinavia.

Currently, the R1a family is defined most broadly by the SNP mutation M420. The recent discovery of M420 resulted in a reorganization of the known family tree of R1a, in particular establishing a new paragroup (designated R1a*) for the relatively rare lineages which are not in the R1a1 branch leading to R1a1a.

R1a and R1a1a are believed to have originated somewhere within Eurasia, most likely in the area from Eastern Europe to South Asia. The most recent studies indicate that South Asia is the most likely region of origin.

Different meanings of "R1a"

The naming system commonly used for R1a remains inconsistent in different published sources, and requires some explanation.

In 2002, the Y chromosome consortium (YCC) proposed a new naming system for haplogroups, which has now become standard. In this system, names with the format "R1" and "R1a" are "phylogenetic" names, aimed at marking positions in a family tree. Names of SNP mutations can also be used to name clades or haplogroups. For example, as M173 is currently the defining mutation of R1, R1 is also R-M173, a "mutational" clade name. When a new branching in a tree is discovered, some phylogenetic names will change, but by definition all mutational names will remain the same.

The widely occurring haplogroup defined by mutation M17 was known by various names, such as "Eu19", in the older naming systems. The 2002 YCC proposal assigned the name R1a to the haplogroup defined by mutation SRY1532.2. This included Eu19 (i.e. R-M17) as a subclade, so Eu19 was named R1a1. The discovery of M420 in 2009 has caused a reassignment of these phylogenetic names. R1a is now defined by the M420 mutation: in this updated tree, the subclade defined by SRY1532.2 has moved from R1a to R1a1, and Eu19 (R-M17) from R1a1 to R1a1a.

Phylogeny (Family Tree)

The R1a family tree now has three major levels of branching, with the largest number of defined subclades within the dominant and best known branch, R1a1a (which, as has been noted, will be found with various names; in particular, as "R1a1" in relatively recent but not the latest literature.)

Roots of R1a

R1a, distinguished by several unique markers including the M420 mutation, is a subclade of haplogroup R1, which is defined by SNP mutation M173. Besides R1a, R1 also has the subclades R1b, defined by the M343 mutation, and the paragroup R1*. There is no simple consensus concerning the places in Eurasia where R1, R1a or R1b evolved.

R1a (R-M420)

R1a, defined by the mutation M420, has two branches: R1a1, defined by the mutation SRY1532.2, which makes up the vast majority; and R1a*, the paragroup, defined as M420 positive but SRY1532.2 negative. (In the 2002 scheme, this SRY1532.2 negative minority was one part of the relatively rare group classified as the paragroup R1*.) Mutations understood to be equivalent to M420 include M449, M511, M513, L62, and L63.

Only isolated samples of the new paragroup R1a* have been found by Underhill et al., mostly in the Middle East and Caucasus: 1/121 Omanis, 2/150 Iranians, 1/164 in the

United Arab Emirates, and 3/612 in Turkey. Testing of 7224 more males in 73 other Eurasian populations showed no sign of this category.

R1a1 (R-SRY1532.2)

R1a1 is currently defined by SRY1532.2, also referred to as SRY10831.2. SNP mutations understood to be always occurring with SRY1532.2 include M448, M459, and M516. This family of lineages is dominated by the very large and well-defined R1a1a branch, which is positive for M17 and M198. The paragroup R1a1* (old R1a*) is positive for the SRY1532.2 marker but lacks either the M17 or M198 markers.

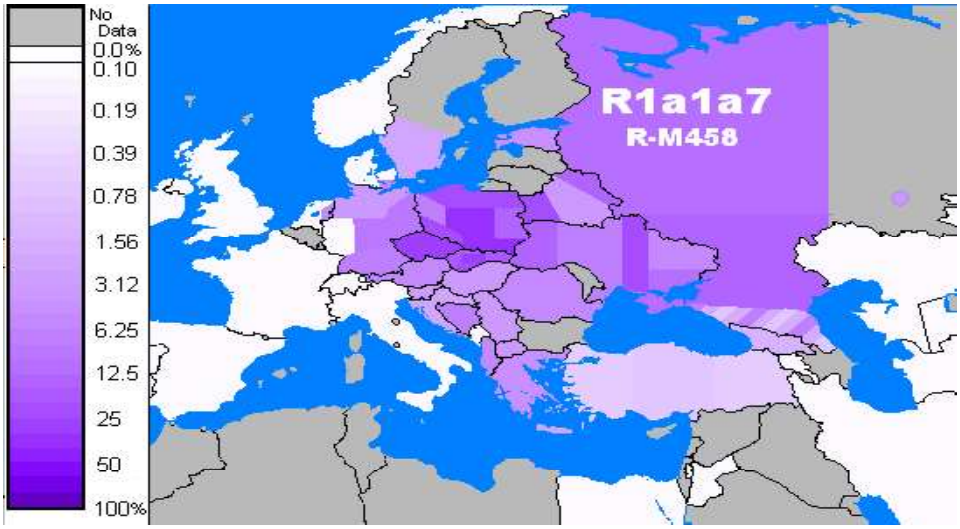
The R1a1* paragroup is apparently less rare than R1* but still relatively unusual, though it has been tested in more than one survey. Underhill et al. for example report 1/51 in Norway, 3/305 in Sweden, 1/57 Greek Macedonians, 1/150 Iranians, 2/734 Ethnic Armenians, and 1/141 Kabardians. While Sahoo et al. reported R1a*(new R1a1*) for 1/15 Himachal Pradesh Rajput samples.

R1a1a (R-M17 or R-M198)

R1a1a (old R1a1) makes up the vast majority of all R1a over its entire geographic range. It is defined by SNP mutations M17 or M198, which have always appeared together in the same men so far. SNP mutations understood to be always occurring with M17 and M198 include M417, M512, M514, M515.

Currently, R1a1a has eight subclades of its own defined by mutations, but the vast majority of the incidence has not yet been categorized and is therefore in the paragroup R1a1a*.

R1a1a subclades



Frequency distribution of R1a1a7 (R-M458)

Currently, of the eight SNP-defined subclades of R1a1a only R1a1a7 has significant frequencies. R1a1a7 is defined by M458 and was found almost entirely in Europe, and with low frequency in Turkey and parts of the Caucasus. Its highest frequencies were found in Central and Southern Poland, particularly near the river valleys flowing northwards to the Baltic sea.

R1a1a7 has its own SNP-defined R1a1a7a subclade, defined by the M334 marker. However this mutation was found only in one Estonian man and may define a very recently founded and small clade.

Relative frequency of R1a1a6 (R-M434) to R1a1a (R-M17)						
Region	People	N	R1a1a-M17		R1a1a6-M434	
			Number	Freq. (%)	Number	Freq. (%)
Pakistan	Baloch	60	9	15%	5	8%
Pakistan	Makrani	60	15	25%	4	7%
Middle East	Oman	121	11	9%	3	2.5%
Pakistan	Sindhi	134	65	49%	2	1%
Table only shows positive sets from N = 3667 derived from 60 Eurasian populations sample, Underhill et al. (2009)						

R1a1a3, defined by the M64.2, M87, and M204 SNP mutations, is apparently rare: it was found in 1 of 117 males typed in southern Iran.

R1a1a6, defined by M434, was detected in 14 people (out of 3667 people tested) all in a restricted geographical range from Pakistan to Oman. This likely reflects a recent mutation event in Pakistan.

R1a1a STR clusters

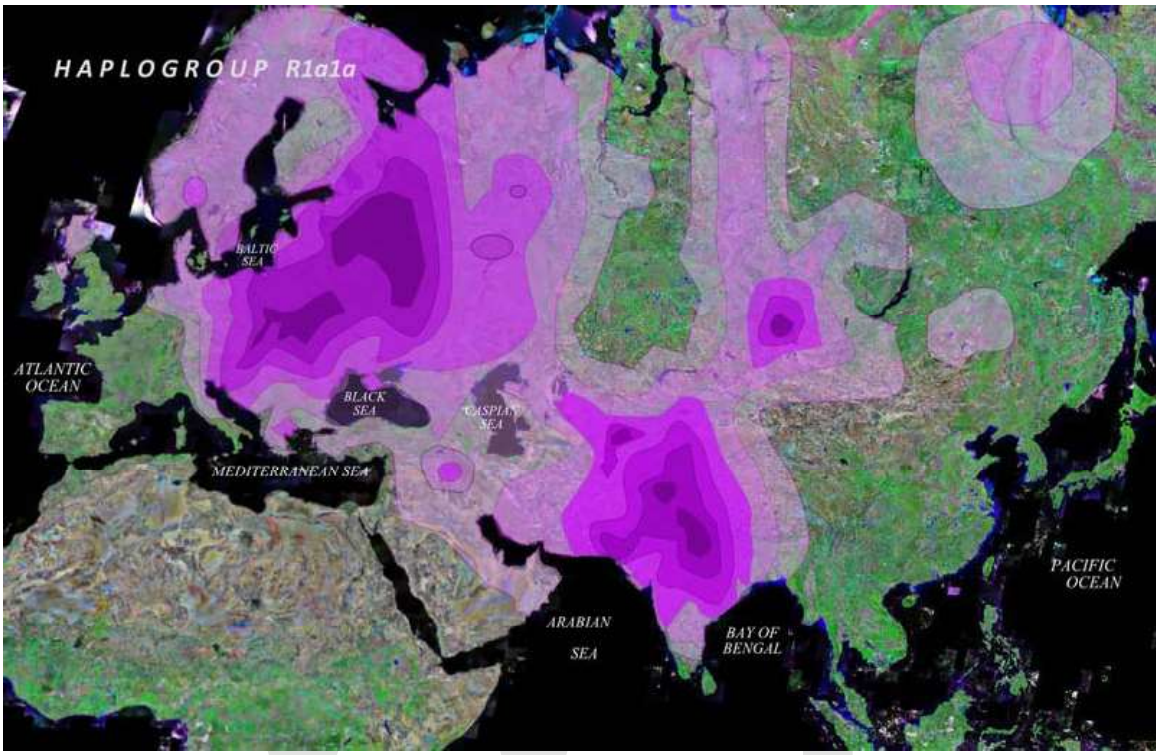
Genetic genealogists looking at high accuracy STR (microsatellite) haplotypes (as used in genealogy) have also identified clusters of similar within R1a1a. Such clusters equate to groups with probable common ancestry, but with no known SNP defining them yet.

Gwozdz (2009) has identified two clusters within R1a1a7 ("P" and "N"). Cluster P was originally identified by Pawlowski (2002) and apparently accounts for about 8% of Polish men, making it the most common clearly identifiable haplotype cluster in Poland. Outside of Poland it is less common. Cluster N is not concentrated in Poland, but is apparently common in many Slavic areas. Gwozdz also identified at least one large cluster of R1a1a* (not having M458), referred to as cluster K. This cluster is common in Poland but not only there.

Klyosov (2009) notes a potential clade identified by a mutation on the relatively stable STR marker DYS388 (to an unusual repeat value of 10, instead of the more common 12), noting that this "is observed in northern and western Europe, mainly in England, Ireland, Norway, and to a much lesser degree in Sweden, Denmark, Netherlands and Germany. In areas further east and south that mutation is practically absent".

Both Gwozdz and Klyosov also note frequent close STR matching between part of the Indian R1a1a population, and part of the Russian and Slavic R1a1a population, indicating apparent links between these populations in a time-frame more recent than the age of R1a1a overall.

Distribution of R1a1a (R-M17 or R-M198)



Frequency distribution of R1a1a, also known as R-M17 and R-M198, adapted from Underhill et al (2009).

R1a has been found in high frequency at both the eastern and western ends of its core range, for example in India and Tajikistan on the one hand, and Poland on the other. Throughout all of these regions, R1a is dominated by the R1a1a (R-M17 or R-M198) sub-clade.

South Asia

In South Asia R1a1a has often been observed with high frequency in a number of demographic groups.

In India, high percentage of this haplogroup is observed in West Bengal Brahmins (72%) to the east, Konkanastha Brahmins (48%) to the west, Khatris (67%) in north and Iyenger Brahmins (31%) of south. It has also been found in several South Indian Dravidian-speaking Adivasis including the Chenchu (26%) and the Valmikis of Andhra Pradesh and the Kallar of Tamil Nadu suggesting that M17 is widespread in Tribal Southern Indians.

Besides these, studies show high percentages in regionally diverse groups such as Manipuris (50%) to the extreme North East and in Punjab (47%) to the extreme North West.

In Pakistan it is found at 71% among the Mohanna of Sindh Province to the south and 46% among the Baltis of Gilgit-Baltistan to the north. While 13% of Sinhalese of Sri Lanka were found to be R1a1a (R-M17) positive.

Hindus of Terai region of Nepal show it at 69%.

Europe

In Europe, R1a, again almost entirely in the R1a1a sub-clade, is found at highest levels among peoples of Eastern European descent (Sorbs, Poles, Russians and Ukrainians; 50 to 65%). In the Baltic countries R1a frequencies decrease from Lithuania (45%) to Estonia (around 30%). Levels in Hungarians have been noted between 20 and 60%

There is a significant presence in peoples of Scandinavian descent, with highest levels in Norway and Iceland, where between 20 and 30% of men are in R1a1a. Vikings and Normans may have also carried the R1a1a lineage westward; accounting for at least part of the small presence in the British Isles.

Haplogroup R1a1a was found at elevated levels amongst a sample of the Israeli population who self-designated themselves as Ashkenazi Jews, originally from European Jewish communities, compared with Sephardic and Middle Eastern Jews. The authors stated that the reasons for these chromosomes in the population is unknown, but could possibly reflect gene flow into Ashkenazi populations from surrounding Eastern European populations, over a course of centuries. This haplogroup finding was apparently consistent with the latest SNP microarray analysis which argued that up to 55 percent of the modern Ashkenazi genome is specifically traceable to Europe.

Ashkenazim were found to have a significantly higher frequency of the R-M17 haplogroup Behar reported R-M17 to be the dominant haplogroup in Ashkenazi Levites (52%), although rare in Ashkenazi Cohanim (1.3%) and Israelites (4%).

In Southern Europe R1a1a is not common amongst the general population, but it is widespread in certain areas. Significant levels have been found in pockets, such as in the Pas Valley in Northern Spain, areas of Venice, and Calabria in Italy. The Balkans shows lower frequencies, and significant variation between areas, for example >30% in Slovenia, Croatia and Greek Macedonia, but <10% in Albania, Kosovo and parts of Greece.

The remains of three individuals, from an archaeological site discovered in 2005 near Eulau (in Saxony-Anhalt, Germany) and dated to about 2600 BCE, tested positive for the Y-SNP marker SRY10831.2. The R1a1 clade was thus present in Europe at least 4600 years ago, and appears associated with the Corded Ware culture.

Central and Northern Asia

R1a1a frequencies vary widely between populations within central and northern parts of Eurasia, but it is found in areas including Western China and Eastern Siberia. This variation is possibly a consequence of population bottlenecks in isolated areas and the movements of Scythians in ancient times and later the Turco-Mongols. High frequencies of R1a1a (R-M17 or R-M198; 50 to 70%) are found among the Ishkashimis, Khojant Tajiks, Kyrgyzs, and in several peoples of Russia's Altai Republic. Although levels are comparatively low amongst some Turkic-speaking groups (*e.g.* Turks, Azeris, Kazakhs, Yakuts), levels are very high in certain Turkic or Mongolic-speaking groups of Northwestern China, such as the Bonan, Dongxiang, Salar, and Uyghurs. R1a1a is also found among certain indigenous Eastern Siberians, including:Kamchatkans and Chukotkans, and peaking in Itel'man at 22%.

Middle East and Caucasus

R1a1a has been found in various forms, in most parts of Western Asia, in widely varying concentrations, from almost no presence in areas such as Jordan, to much higher levels in parts of Turkey and Iran.

Wells et al. (2001), noted that in the western part of the country, Iranians show low R1a1a levels, while males of eastern parts of Iran carried up to 35% R1a. Nasidze et al. (2004) found R1a in approximately 20% of Iranian males from the cities of Tehran and Isfahan. Regueiro et al. (2006), in a study of Iran, noted much higher frequencies in the south than the north.

Turkey also shows high but unevenly distributed R1a levels amongst some sub-populations. For example Nasidze et al. (2005) found relatively high levels amongst Kurds (12%) and Zazas (26%).

Further to the north of these Middle Eastern regions on the other hand, R1a levels start to increase in the Caucasus, once again in an uneven way. Several populations studied have shown no sign of R1a, while highest levels so far discovered in the region appears to belong to speakers of the Karachay-Balkar language amongst whom about one quarter of men tested so far are in haplogroup R1a1a.

Origins and hypothesized migrations of R1a1a

Most discussions purportedly of R1a origins are actually about the origins of the dominant R1a1a (R-M17 or R-M198) sub-clade. Data so far collected indicates that there two widely separated areas of high frequency, one in South Asia, around Indo-Gangetic Plain, and the other in Eastern Europe, around Poland and Ukraine. The historical and prehistoric possible reasons for this are the subject of on-going discussion and attention amongst population geneticists and genetic genealogists, and are considered to be of potential interest to linguists and archaeologists also.

In 2009, several large studies of both old and new STR data concluded that while these two separate "poles of the expansion" are of similar age, South Asian R1a1a is apparently older than Eastern European R1a1a, suggesting that South Asia is the more likely locus of origin.

South Asian origin hypothesis

An increasing number of studies have found South Asia to have the highest level of diversity of Y-STR haplotype variation within R1a1a. On this basis, while several studies have concluded that the data is consistent with South Asia as the likely original point of dispersal (for example, Kivisild et al. (2003), Mirabal et al. (2009) and Underhill et all. (2009)) a few have actively argued for this scenario (for example Sengupta et al. (2005), Sahoo et al. (2006), Sharma et al. (2009). A survey study as of December 2009, including a collation of retested Y-DNA from previous studies, makes a South Asian R1a1a origin the strongest proposal amongst the various possibilities.

Central Asia

Cordaux et al. (2004) argued, citing data from 3 earlier publications, that R-M17 (R1a1a) Y chromosomes most probably have a central Asian origin. Central Asia is still considered a possible place of origin by Mirabal et al. (2009) after their larger analysis of more recent data. However these authors also consider other parts of Asia, particularly South Asia, to likely places of origin.

Middle East

As mentioned above, R1a haplotypes are less common in most of the Middle East than they are in either South Asia or Eastern Europe or much of Central Asia. It has nevertheless been mentioned in speculation about the origins of the clade. This is both because there are above-described pockets of high frequency and diversity, for example in some parts of Iran and amongst some Kurdish populations. A Middle Eastern origin for R1a has long been considered a possibility, and is still considered to be consistent with known data.

Eastern European migration hypotheses

Coalescent time estimates for R1a1a(xM458) STR from Underhill et al. (2009)

Location	T_D
W. India	15,800
Pakistan	15,000
Nepal	14,200
India	14,000
Oman	12,500
N. India	12,400
S. India	12,400
Caucasus	12,200

E. India	11,800
Poland	11,300
Slovakia	11,200
Crete	11,200
Germany	9,900
Denmark	9,700
UAE	9,700

A widely cited theory proposed in 2000 that there may have been two expansions: first, R1a1a originally spreading from a Ukrainian refugium during the Late Glacial Maximum; and then, the spread being magnified by the expansion of males from the Kurgan culture. A recent survey argues that R1a1a could be old enough for this scenario, but find it more likely that it was initially in Asia even if it was in parts of Europe by approximately 11,000 years ago.

Most age estimates for R1a1a having such an early presence in Europe come from papers using the "evolutionarily effective" methodology described by Zhivotovsky et al. (2004), the latest such example being Mirabal et al. (2009) and Underhill et al. (2009). Researchers using this dating method therefore conclude that any Neolithic or more recent dispersals of R1a1a do not represent the initial spread of the whole clade, and might be more visible in the distribution of a subclade or subclades. Underhill et al. (2009) remark on the "geographic concordance of the R1a1a7-M458 distribution with the Chalcolithic and Early Bronze Age Corded Ware (CW) cultures of Europe". However they also note evidence contrary to a connection: Corded Ware period human remains at Eulau from which Y-DNA was extracted of R1a haplogroup appear to be R1a1a*(xM458) (which they found most similar to the modern German R1a1a* haplotype.)

In papers where the Zhivitovsky method is not the only method used, Europe's R1a1a diversity is generally understood to have been shaped more significantly by more recent events, including not only the Bronze Age, but also the spread of Slavic languages. Dupuy et al. (2005) speculated that "R1a [in Norway] might represent the spread of the Corded Ware and Battle-Axe cultures from central and east Europe." Luca et al. (2006), looking at data from the Czech Republic suggested there was evidence for a rapid demographic expansion approximately 1500 years ago. Rebala et al. (2007) also detected Y-STR evidence of a recent Slavic expansion from the area of modern Ukraine. Gwodzdz (2009) saw evidence for a "rapid population expansion somewhat less than 1,500 years ago in the area that is now Poland".

Steppe cultures

Archaeologists recognize a complex of inter-related and relatively mobile cultures living on the Eurasian steppe, part of which protrudes into Europe as far west as Ukraine. These cultures from the late Neolithic and into the Iron Age, with specific traits such as Kurgan burials and horse domestication, have been associated with the dispersal of Indo-European languages across Eurasia. Nearly all samples from Bronze and Iron Age graves

in the Krasnoyarsk area in south Siberia belonged to R1a1-M17 and appeared to represent an eastward migration from Europe.

Geneticists believing that they see evidence of R1a1a gene-flow from the Eurasian Steppe to India have frequently proposed the involvement of these Steppe cultures in the process. Such a Steppe origin for all or part R1a1a continues to be argued on the basis of DNA results from ancient remains from several South Siberian late Kurgan sites, including some from the Andronovo culture. However, in recent discussions of this theory it is considered only to apply to a part of R1a1a, making this theory no longer incompatible with other origins theories for R1a more broadly defined.

Chapter- 8

Genealogical DNA Test

A **genealogical DNA test** examines the nucleotides at specific locations on a person's DNA for genetic genealogy purposes. The test results are not meant to have any informative medical value and do not determine specific genetic diseases or disorders; they are intended only to give genealogical information. Genealogical DNA tests generally involve comparing the results of living individuals to historic populations.

Procedure

The general procedure for taking a genealogical DNA test involves taking a painless cheek-scraping (also known as a buccal swab) at home and mailing the sample to a genetic genealogy laboratory for testing. Some laboratories use mouth wash or chewing gum instead of cheek swabs. Some laboratories, such as the Human Origins Genotyping Laboratory (HOGL) at the University of Arizona, offer to store DNA samples for ease of future testing. All United States laboratories will destroy the DNA sample upon request by the customer, guaranteeing that a sample is not available for further analysis.

Types of tests

The most popular ancestry tests are Y chromosome (Y-DNA) testing and mitochondrial DNA (mtDNA) testing which test direct-line paternal and maternal ancestry, respectively. DNA tests for other purposes attempt, for example, to determine a person's comprehensive genetic make-up and/or ethnic origins.

Y chromosome (Y-DNA) testing

A man's patrilineal ancestry, or male-line ancestry, can be traced using the DNA on his Y chromosome (Y-DNA) through Y-STR testing. This is useful because the Y chromosome passes down almost unchanged from father to son, ie, the non recombining and sex

determining regions of the Y chromosome do not change. A man's test results are compared to another man's results to determine the time frame in which the two individuals shared a most recent common ancestor or MRCA. If their test results are a perfect, or nearly perfect match, they are related within genealogy's time frame.

Each person can then look at the other's father-line information, typically the names of each patrilineal ancestor and his spouse, together with the dates and places of their marriage and of both spouses' births and deaths. This information table will be referred to again within the mtDNA testing section below as the (matrilineal) "information table". The two matched persons may find a common ancestor or MRCA, as well as whatever information the other already has about their joint patrilineal ancestry prior to the MRCA—which might be a big help to one of them. Or if not, both keep trying to extend their patrilineal ancestry further back in time. Each may choose to have their test results included in their surname's "Surname DNA project". And each receives the other's contact information if the other chose to allow this. They may correspond, and may work together in the future on joint research.

Women who wish to determine their direct paternal DNA ancestry can ask their father, brother, paternal uncle, paternal grandfather, or a cousin who shares a common patrilineal ancestry (the same Y-DNA) to take a test for them.

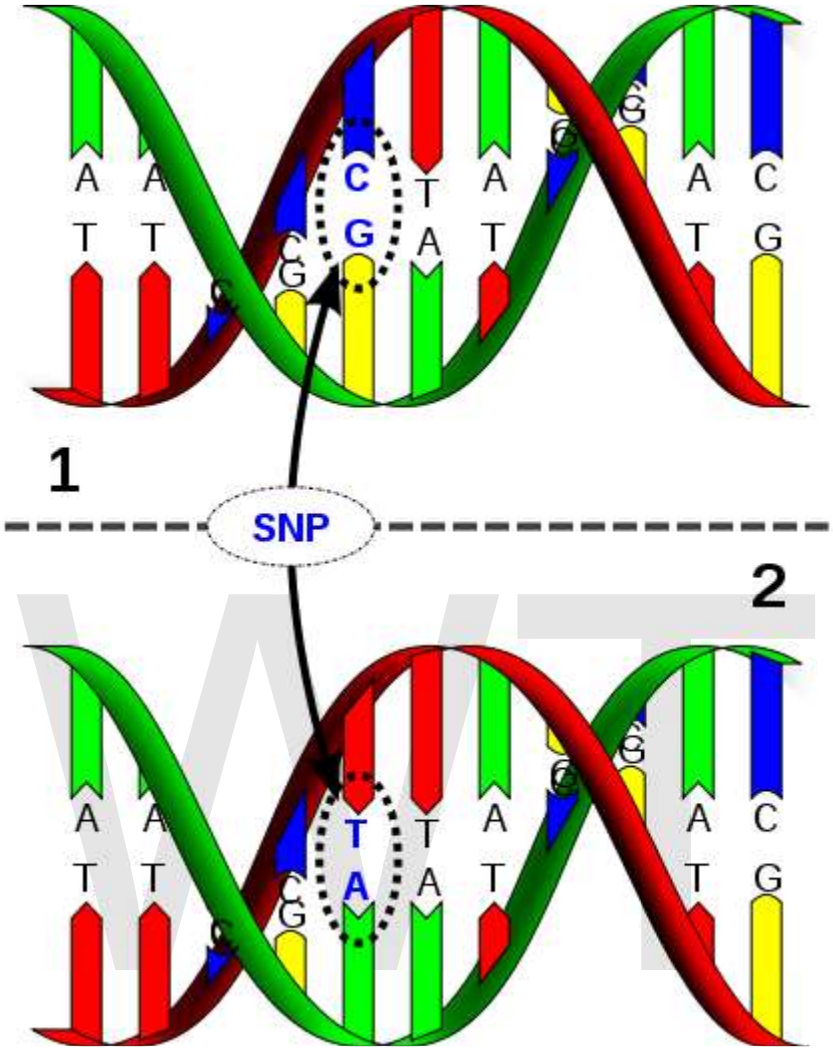
What gets tested

Y-DNA testing involves looking at STR segments of DNA on the Y chromosome. The STR segments which are examined are referred to as genetic markers and occur in what is considered "junk" DNA.

STR markers

A chromosome contains sequences of repeating nucleotides known as short tandem repeats (STRs). The number of repetitions varies from one person to another and a particular number of repetitions is known as an allele of the marker. An STR on the Y chromosome is designated by a **DYS** number (**D**NA **Y**-chromosome **S**egment number). The example below shows the allele of Rumpelstiltskin's DYS393 marker is 12, also called the marker's "value". The value 12 means the DYS393 sequence of nucleotides is repeated 12 times—with a DNA sequence of $(AGAT)_{12}$.

SNP markers



Strand 1 differs from strand 2 at a single base pair location (a C → T polymorphism).

A single-nucleotide polymorphism (SNP) is a change to a single nucleotide in a DNA sequence. The relative mutation rate for an SNP is extremely low. This makes them ideal for marking the history of the human genetic tree. SNPs are named with a letter code and a number. The letter indicates the lab or research team that discovered the SNP. The number indicates the order in which it was discovered. For example, M173 is the 173rd SNP documented by the Human Population Genetics Laboratory at Stanford University, which uses the letter M.

Understanding test results

Y-DNA tests generally examine 10-67 STR markers on the Y chromosome, but over 100 markers are available. STR test results provide the personal haplotype. SNP results indicate the haplogroup.

Haplotype

A Y-DNA haplotype is the numbered results of a genealogical Y-DNA test. Each allele value has a distinctive frequency within a population. For example, at DYS455, the results will show 8, 9, 10, 11 or 12 repeats, with 11 being most common. For high marker tests the allele frequencies provide a signature for a surname lineage.

Kit	Surname	Haplo	393	390	19	391	385a	385b	426	388	439	389-1	392	389-2	458	459a	459b	455	454	447	437	448	449	464a	464b	464c	464d
11111	Rumpelstiltskin	Q	12	23	13	10	16	17	12	12	13	14	14	31	18	8	9	11	11	27	13	19	28	14	14	15	15

The test results are then compared to another project member's results to determine the time frame in which the two people shared a most recent common ancestor (MRCA). If the two tests match perfectly on 37 markers, there is a 50% probability that the MRCA was fewer than 2 to 3 generations ago, 90% probability that the MRCA was fewer than 5 generations ago, and 95% probability that the MRCA was fewer than 7 generations ago.

Kit	Surname	Haplo	393	390	19	391	385a	385b	426	388	439	389-1	392	389-2	458	459a	459b	455	454	447	437	448	449	464a	464b	464c	464d
11111	Rumpelstiltskin	Q	12	23	13	10	16	17	12	12	13	14	14	31	18	8	9	11	11	27	13	19	28	14	14	15	15
11178	Rumpelstiltskin	Q	12	23	13	10	16	17	12	12	13	14	14	31	18	8	9	11	11	27	13	19	28	14	14	15	15

Before choosing a test, it is important for an individual to check the number of markers that will be tested. For example, the Genographic Project looks at only 12 markers, while most laboratories and surname projects recommend testing at least 25. The more markers that are tested, the more discriminating and powerful the results will be. A 12-marker STR test is usually not discriminating enough to provide conclusive results for a common surname.

STRs results may also indicate a likely haplogroup, though this can only be confirmed by specifically testing for that Haplogroups' single-nucleotide polymorphisms (SNPs).

Haplogroup prediction

A person's haplogroup can often be inferred from their haplotype, but can be proven only with a Y-chromosome SNP tests (Y-SNP test). In addition, some companies offer sub-clade tests, such as for Haplogroup G. For example, Haplogroup G has a known modal haplotype:

DYS markers	385a	385b	388	389i	389ii	390	391	392	393	394	426	437	439	444	447	448	449	454	455	458	459a	459b	464a	464b	464c	464d
Haplogroup G: Modal STR values	14	14	12	12	29	22	10	11	14	15	11	16	11	23	21	31	11	11	16	9	9	12	13	13	14	

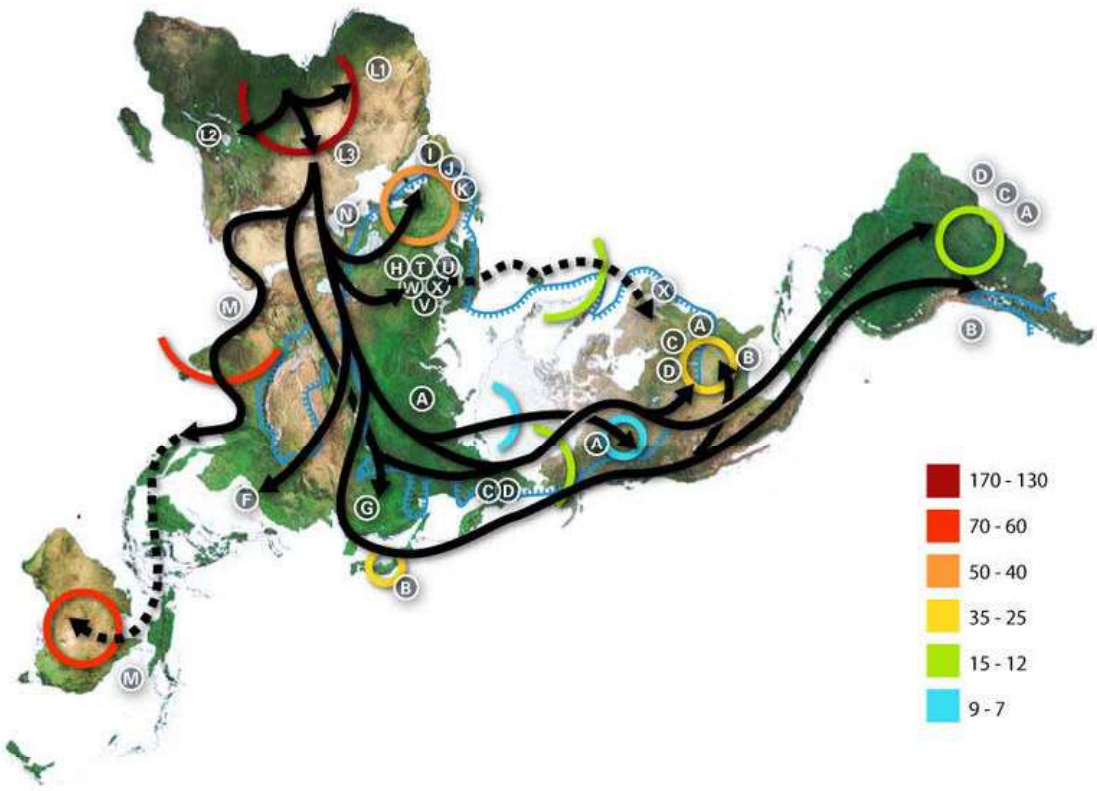
Few haplotypes will exactly match the modal values for Haplogroup G. One can consult an allele frequency table to determine the likelihood of remaining in Haplogroup G based on the variations observed.

Additional predictions include:

- If DYS426 is 12 and DYS392 is 11, one is probably a member of haplogroup R1a1.
- If DYS426 is 12 and DYS392 is not 11, one is probably a member of haplogroup R1b.
- If DYS426 is 11, one is probably a member of haplogroup G,I, or J.
- If DYS426 is 11 and DYS388 is 12, one is in the known modal haplotype for G shown above.

A Bayes classifier to predict the haplogroup probabilities for an observed haplotype is available on the web: Whit Athey Haplogroup Predictor.

Mitochondrial DNA (mtDNA) testing



Map of human migration out of Africa, according to Mitochondrial DNA. The numbers represent thousands of years before present time. The blue line represents the area covered in ice or tundra during the last great ice age. The North Pole is at the center. Africa, the center of the start of the migration, is at the top left and South America is at the far right.

A person's matrilineal or mother-line ancestry can be traced using the DNA in his or her mitochondria, the mtDNA, as follows: This mtDNA is passed down by the mother unchanged, to all children. If a perfect match is found to another person's mtDNA test results, one may find a common ancestor in the other relative's (matrilineal) "information table", similar to the patrilineal or Y-DNA testing case above. However, because mtDNA mutations are very rare, a *nearly* perfect match is not as helpful as it is for the above patrilineal case. In the matrilineal case, it takes a perfect match to be very helpful.

Note that, in cultures lacking *matrilineal* surnames to pass down, neither relative above is likely to have as many generations of ancestors in their matrilineal information table as in the above patrilineal or Y-DNA case.

Some people cite paternal mtDNA transmission as invalidating mtDNA testing, but this has not been found problematic in genealogical DNA testing, nor in scholarly population genetics studies.

What gets tested

mtDNA by current conventions is divided into three regions. They are the coding region (00577-16023) and two Hyper Variable Regions (HVR1 [16024-16569], and HVR2 [00001-00576]). All test results are compared to the mtDNA of a European in Haplogroup H2a2a. This early sample is known as the Cambridge Reference Sequence (CRS). A list of single nucleotide polymorphisms (SNPs) is returned. The relatively few "mutations" or "transitions" that are found are then reported simply as differences from the CRS, such as in the examples just below.

The two most common mtDNA tests are a sequence of HVR1 and a sequence of both HVR1 and HVR2. Some mtDNA tests may only analyze a partial range in these regions. Some people are now choosing to have a full sequence performed, to maximize their genealogical help. The full sequence is still somewhat controversial because it may reveal medical information.

Understanding test results

The most basic of mtDNA tests will sequence Hyper Variable Region 1 (HVR1). HVR1 nucleotides are numbered 16024-16569. Some test reports might omit the 16 prefix from HVR1 results, i.e. 519C and not 16519C.

Region	HVR1	HVR2
Differences from CRS	111T,223T,259T,290T,319A,362C	Not Tested

More extensive tests will also sequence Hyper Variable Region 2 (HVR2). HVR2 nucleotides are numbered 00001-00576.

Region	HVR1	HVR2
Differences from CRS	111T,223T,259T,290T,319A,362C	073G,146C,153G

Geographic origin tests

Autosomal tests that test the recombining chromosomes are available. These attempt to measure an individual's mixed geographic heritage by identifying particular markers, called ancestry informative markers or AIM, that are associated with populations of specific geographical areas. The tests' validity and reliability have been called into question but they continue to be popular. Anomalous findings most often result from databases too small to associate markers with all the areas where they occur in indigenous populations.

Biogeographical ancestry

Autosomal DNA testing purports either to determine the "genetic percentages" of a person's ancestry from particular continents/regions or to identify the countries and

"tribes" of origin on an overall basis. *Admixture* tests arrive at these percentages by examining SNPs, which are locations on the DNA where one nucleotide has "mutated" or "switched" to a different nucleotide. Tests' listing geographical places of origin use alleles—individual and family variations on various chromosomes across the genome analyzed with the aid of population databases. As further detailed below, this latter type of test concentrates on standard identity markers, such as the CODIS profile, combined with databases such as OmniPop, ENFSI and proprietary adaptations of published studies.

The *admixture* tests are designed to tell what percentages a person has of ancestry of Native American, "European", East Asian, and Sub-Saharan African. One company describes these four biogeographic groups as follows:

- Native American: Populations that migrated from Asia to inhabit North, South and Central America.
- European: European, Middle Eastern and South Asian populations from the Indian subcontinent, including India, Pakistan and Sri Lanka.
- East Asian: Japanese, Chinese, Mongolian, Korean, Southeast Asian and Pacific Islander populations, including populations native to the Philippines.
- African: Populations from Sub-Saharan Africa such as Nigeria and Congo region.

Based on customer feedback, the company in June 2007 introduced a new version of its EURO DNA test, with a more limited range of countries, that promises to provide more meaningful clues to one's European ancestry. Both tests: the four-part ethnicity estimate and EURO DNA test, identify a high number of so-called Ancestry Informative Markers (AIM), whose genetic distance between populations reflects the populations' geographic distance from each other. The location and variation of these AIMs are proprietary to the company and have never been published.

In 2006, another company developed an autosomal DNA ancestry-tracing product that combined the traditional CODIS markers used by law enforcement officers and the judicial system with OmniPop, a population database developed by San Diego detective Brian Burritt. Customers received matches to their profile's frequency of occurrence in world populations, as well as a breakout for European ancestry based on the European Network of Forensic Science Institutes (ENFSI). As a public service, the company has supported the expansion of OmniPop, which currently encompasses over 360 populations, double that of its first release. The ENFSI calculator uses data from 24 European populations (5700 profiles). The two databases must be searched separately, because they are based on two different sets of markers. The company sells its product as the DNA Fingerprint Test. The 16 markers incorporated in its results are: D8S1179, D21S11, D7S820, CSFIPO, D3S1358, THO1, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818, and FGA.

The theory behind using a forensic profile for ancestry tracing is that the alleles' respective frequency of occurrence develops over generations with equal input of the two parents, since for each location we take one value from our mother and one from our

father. It thus serves as a window into a person's total ancestral composition. The configuration of scores reflects inherited changes from all previous generations in all ancestral lines, and can predict an individual's unique probable ethnic matches based on the profile's frequency or rarity in different populations.

To give an idea of the inclusiveness of the latest version of OmniPop, the following are the last populations that have been added:

- Greek
- Sikkim (India)
- Bhutia (India)
- Italian
- Argentinian (Misiones)
- Hungarian (E. Romani)
- Hungarian (Ashkenazim)
- Romanian (Szekler)
- Romanian (Csango)
- Tibet (Luoba)

As studies from more populations are included, the accuracy of results should improve, leading to a more informative picture of one's ancestry.

Along the same lines, yet another company identifies the indigenous and diaspora populations in which an individual's autosomal STR profile is most common. This test examines autosomal STRs, which are locations on a chromosome where a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other. The populations in which the individual's profile is most common are identified and assigned a likelihood score. The individual's profile is assigned a likelihood of membership in each of thirty-four world regions:

- Caucasian
    - European:
        - Eastern European: The Slavic-speaking region of eastern Europe.
        - Finno-Ugrian: The Uralic-speaking region of northeastern Europe.
        - Mediterranean: The Romance-speaking region of southwestern Europe.
        - Northwest European: The Celtic and Germanic-speaking region of northwestern Europe.
        - Aegean: Anatolia region, modern territories of Southern Italy and Sicily, Greece, and Turkey.
    - Near Eastern
        - Arabian: The Arabian Peninsula.
        - North African: Populations of the Atlas Mountains and Sahara Desert.
        - Mesopotamian: The historical “Cradle of Western Civilization”, including modern Iran, Iraq and nearby territories.

        - Levantine: Populations along the coast of the eastern Mediterranean Sea.

- South Asian:
    - Eastern India
    - North India
    - South India

- East Asian:

- Subsaharan African:
    - East African
    - Southern African
    - West African

- American Indian

- Polynesian

The STR analysis measures the frequency of a person's DNA profile within major world regions. Unlike SNP admixture tests, this analysis is based on objectively identified world regions and does not depend on any system of presumed biogeographic classifications. As most STR analysis examines markers chosen for their high intra-group variation, the utility of these particular STR markers to access inter-group relationships may be greatly diminished.

United States

Because of its history of immigration, slavery, and significant indigenous peoples, people of the United States have been interested in using genealogical DNA studies to help them learn more about their ancestry.

United States - Native American ancestry

Autosomal testing, Y-DNA, and mtDNA testing can be conducted to determine Amerindian ancestry. A mitochondrial Haplogroup determination test based on mutations in Hypervariable Region 1 and 2 may establish whether a person's direct female line belongs to one of the canonical Native American Haplogroups, A, B, C, D or X. If one's DNA belonged to one of those groups, the implication would be that he or she is, in whole or part, Native American.

As political entities, tribes have established their own requirements for membership, often based on at least one of a person's ancestors having been included on tribal-specific Native American censuses (or final rolls) prepared during treaty-making, relocation to reservations or apportionment of land in the late 19th century and early 20th century. One example is the Dawes Rolls. In addition, the U.S. government does not consider DNA as

admissible evidence for enrollment in any federally recognized tribe or reception of benefits. Tribes are political constructs, not genetic populations.

The vast majority of Native American individuals do belong to one of the five identified mtDNA Haplogroups. Many Americans are just discovering they have some percentage of Native ancestry. Some attempt to validate their heritage with the goal of gaining admittance into a tribe, but most tribes do not use DNA results in that way. These tests may be useful for adoptees to discover Native American ancestry.

United States - African ancestry

Y-DNA and mtDNA testing may be able to determine with which peoples in present-day African country a person shares a direct line of part of his or her ancestry, but patterns of historic migration and historical events cloud the tracing of ancestral groups. Testing company African Ancestry maintains an "African Lineage Database" of African lineages from 30 countries and over 160 ethnic groups. Due to joint long histories in the US, approximately 30% of African American males have a European Y chromosome haplogroup Approximately 58% of African Americans have the equivalent of one great-grandparent (12.5 percent) of European ancestry. Only about 5% have the equivalent of one great-grandparent of Native American ancestry. By the early 19th century, substantial families of Free Persons of Color had been established in the Chesapeake Bay area who were descended from people free during the colonial period; most of those have been documented as descended from white women (servant or free) and African men (servant, slave or free). Over time various groups married more within mixed-race, black or white communities.

According to authorities like Salas, nearly three-quarters of the ancestors of African Americans taken in slavery came from regions of West Africa. The African-American movement to discover and identify with ancestral tribes has burgeoned since DNA testing became available. Often members of African-American churches take the test as groups. African Americans cannot easily trace their ancestry during the years of slavery through surname research, census and property records, and other traditional means. Genealogical DNA testing may provide a tie to regional African heritage.

United States - Melungeon testing

Melungeona are one of numerous multiracial groups in the United States with origins wrapped in myth. The historical research of Paul Heinegg has documented that many of the groups in the Upper South were descended from mixed-race people who were free in colonial Virginia and descended from unions between the Europeans and Africans. They moved to the frontiers of Virginia, North Carolina, Kentucky and Tennessee to gain some freedom from the racial barriers of the plantation areas. Several efforts, including a number of ongoing studies, have examined the genetic makeup of families historically identified as Melungeon. Most results point primarily to a mixture of European and African, which is supported by historical documentation. Some may have a very small amount of Native American lineages (none in one study). Though some companies

provide additional Melungeon research materials with Y-DNA and mtDNA tests, any test will allow comparisons with the results of current and past Melungeon DNA studies.

General interest

Cohanim ancestry

The Cohanim (or Kohanim) is a patrilineal priestly line of descent in Judaism. According to the Bible, the ancestor of the Cohanim is Aaron, brother of Moses. Many believe that descent from Aaron is verifiable with a Y-DNA test: the first published study in genealogical Y chromosome DNA testing found that a significant percentage of Cohens had distinctively similar DNA, rather more so than general Jewish or Middle Eastern populations. These Cohens tended to belong to Haplogroup J, with Y-STR values clustered unusually closely around a haplotype known as the Cohen Modal Haplotype (CMH). This could be consistent with a shared common ancestor, or with the hereditary priesthood having originally been founded from members of a single closely related clan.

But, the original studies tested only six Y-STR markers, which is considered a low-resolution test. Such a test does not have the resolution to prove relatedness, nor to estimate reliably the time to a common ancestor. The Cohen Modal Haplotype (CMH), while notably frequent among Cohens, also appears in the general populations of haplogroups J1 and J2 with no particular link to the Cohen ancestry. So while many Cohens have haplotypes close to the CMH, many more of such haplotypes worldwide belong to people with no likely Cohen connection at all. According to researchers (Hammer), it is only the CMH that is found in J1 that is to be attributed to the Aaron lineage, not the CMH in J2. Jews with the CMH in both J1 and J2 cannot all be descended from one man who lived approximately 3,300 years ago, because J1 diverged from J2 10,000 years ago.

Resolution may be increased by the testing of more than six Y-STR markers. For some, this could help to establish relatedness to particular recent Cohen clusters. For many, the testing is unlikely to distinguish definitively shared Cohen ancestry from that of the more general population distribution. So far no published research indicates what extended Y-STR haplotype distributions appear to be characteristic of Cohens.

Although some high-resolution testing has been done, to date the results have not been released.

European testing

For people with European maternal ancestry, mtDNA tests are offered to determine which of eight European maternal "clans" the direct-line maternal ancestor belonged to. This mtDNA haplotype test was popularized in the book *The Seven Daughters of Eve*.

SNP testing may enable mostly European individuals to determine to which Sub-European population they belong:

- Northern European subgroup (NOR) - mostly Northern and Southwestern European
- Southeastern European (Mediterranean) subgroup (MED) - mostly Southeastern Europeans (Greeks or Turks)
- Middle Eastern subgroup (MIDEAS) - mostly Middle Eastern
- South Asian subgroup (SA) - mostly South Asian from the Indian sub-continent (i.e. Indian)

Hindu testing

The 49 established *gotras* are clans or families whose members trace their descent to a common ancestor, usually a sage of ancient times. The gotra proclaims a person's identity and a "gotraspeak" is required to be presented at Hindu ceremonies. People of the same gotra are not allowed to marry.

One company says it can use a 37-marker Y-DNA test to "verify genetic relatedness and historical gotra genealogies for Hindu and Buddhist engagements, marriages and business partnerships." This has not been supported by independent research. Any Y-DNA test can be used to compare results with another person whose gotra is known.

Benefits

Genealogical DNA tests have become popular due to the ease of testing at home and their supplementing genealogical research. Genealogical DNA tests allow for an individual to determine with high accuracy whether he or she is related to another person within a certain time frame, or with certainty that he or she is not related. DNA tests are perceived as more scientific, conclusive and expeditious than searching the civil records. But, they are limited by restrictions on lines which may be studied. The civil records are always only as accurate as the individuals who provided or wrote the information.

The aforementioned Y-DNA testing results are normally stated as probabilities: For example, a perfect 12/12 marker test match gives a 90% likelihood of the most recent common ancestor (MRCA) being within 23 generations, while a 67 of 67 marker match gives the same 90% likelihood of the MRCA being within 4 generations back.

As presented above in mtDNA testing, if a perfect match is found, the mtDNA test results can be helpful. In some cases, research according to traditional genealogy methods encounters difficulties due to the lack of regularly recorded matrilineal surname information in many cultures.

Drawbacks

Common concerns about genealogical DNA test are cost and privacy issues (some testing companies retain samples and results for their own use without a privacy agreement with subjects). The most common complaint from DNA test customers is the failure of the company to make results understandable to them.

DNA tests can do some things well, but there are constraints. Testing of the Y-DNA lineage from father to son may reveal complications, due to unusual mutations, secret adoptions, and false paternity (i.e. the father in one generation is not the father in birth records.) According to some genomics experts, autosomal tests may have a margin of error up to 15% and blind spots.

Some users have recommended that there be government or other regulation of ancestry testing to ensure more standardization.

Medical information

Though genealogical DNA test results generally have no informative medical value and are not intended to determine genetic diseases or disorders, a correlation exists between a lack of DYS464 markers and infertility, and between mtDNA haplogroup H and protection from sepsis. Certain haplogroups have been linked to longevity.

The testing of full mtDNA sequences is still somewhat controversial as it may reveal medical information. The field of linkage disequilibrium, unequal association of genetic disorders with a certain mitochondrial lineage, is in its infancy, but those mitochondrial mutations that have been linked are searchable in the genome database Mitomap. The National Human Genome Research Institute operates the Genetic And Rare Disease Information Center that can assist consumers in identifying an appropriate screening test and help locate a nearby medical center that offers such.

DNA in genealogy software

Some genealogy software programs now allow recording DNA marker test results, allowing for tracking of both Y-chromosome and mtDNA tests, and recording results for relatives. DNA-family tree wall charts are available.