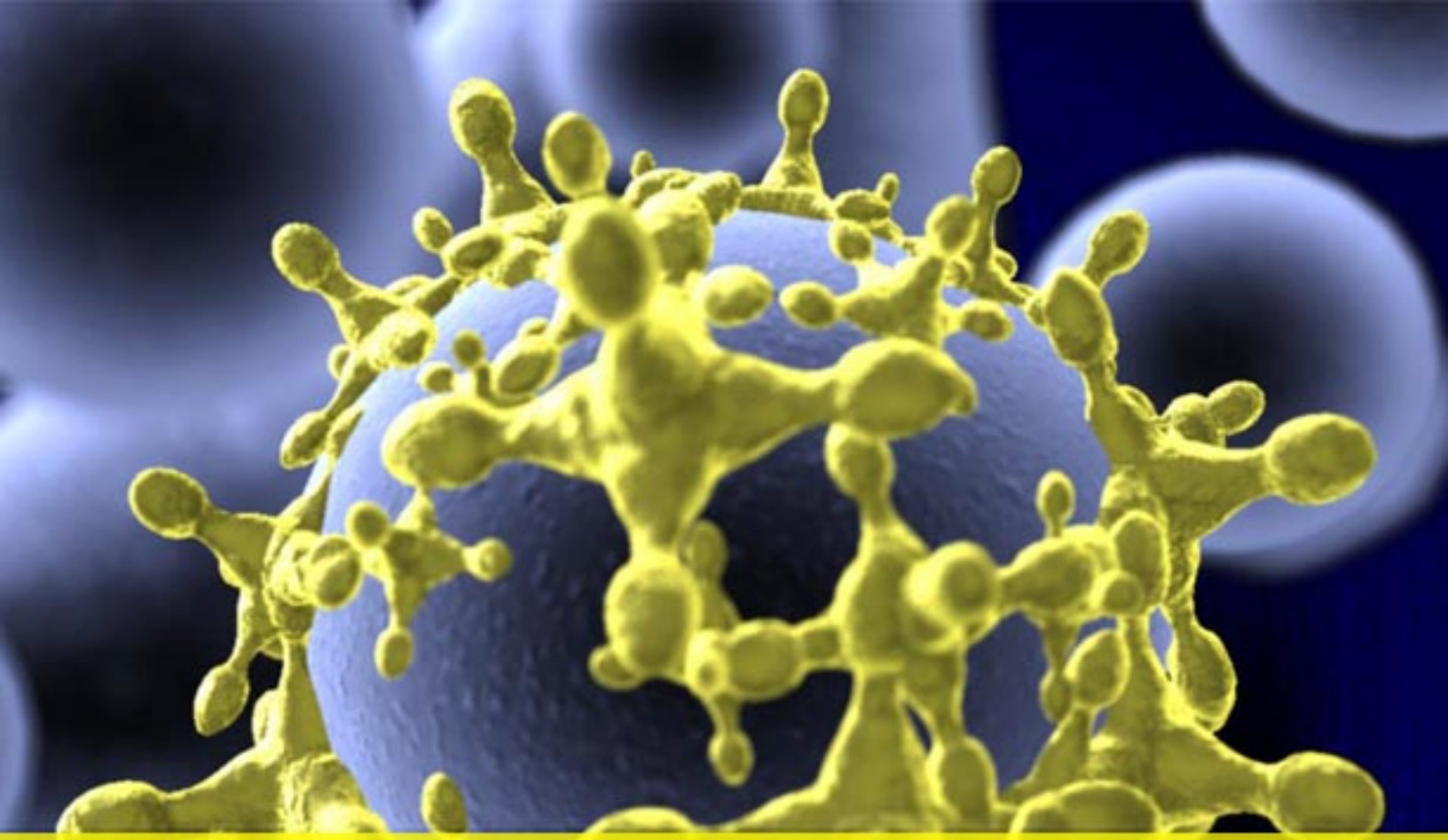


Genetic Genealogy

Serenity Oswald

First Edition, 2012

ISBN 978-81-323-3149-0

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

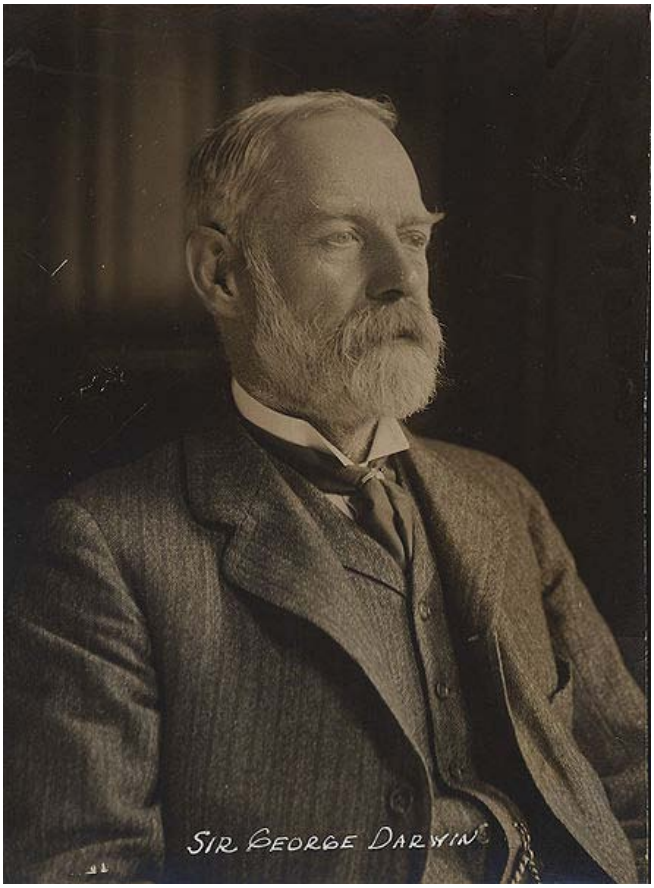
Table of Contents

Chapter- 1

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.

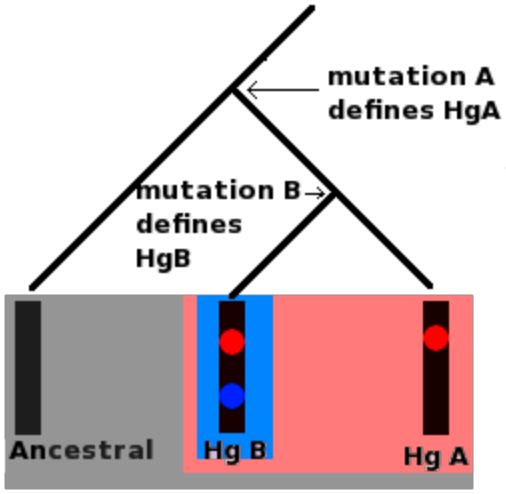
Typical customers and interest groups

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- 2

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.

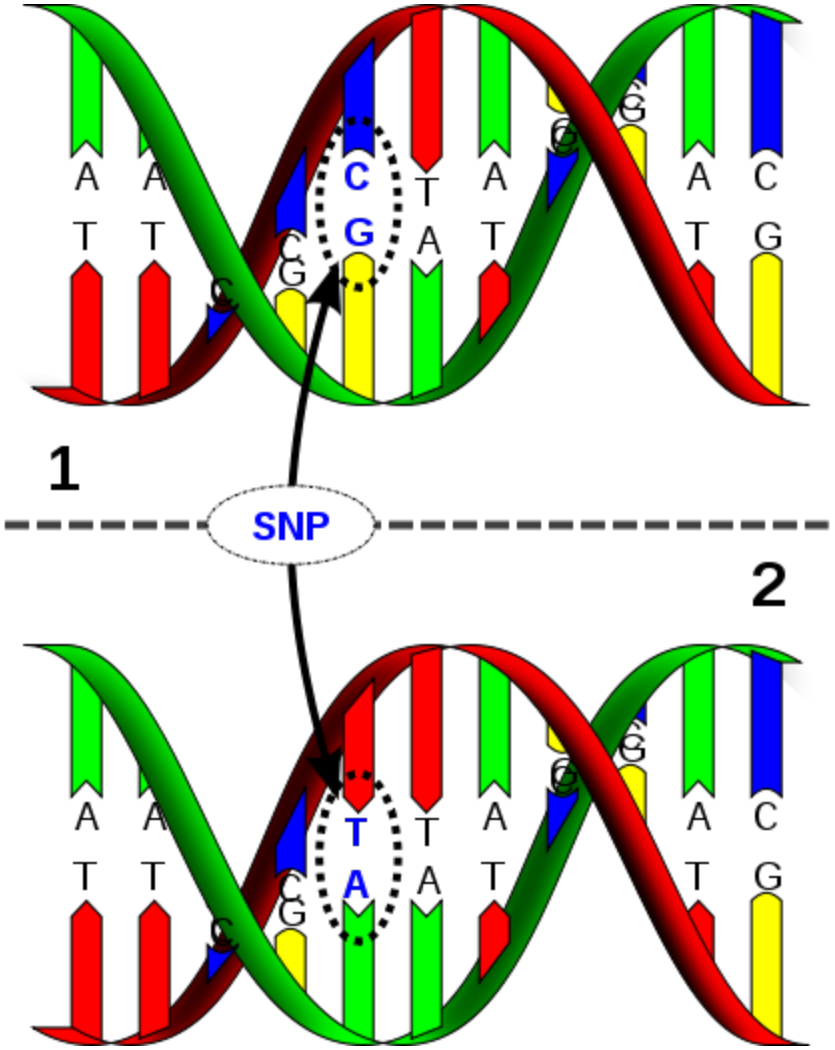
What gets tested

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	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

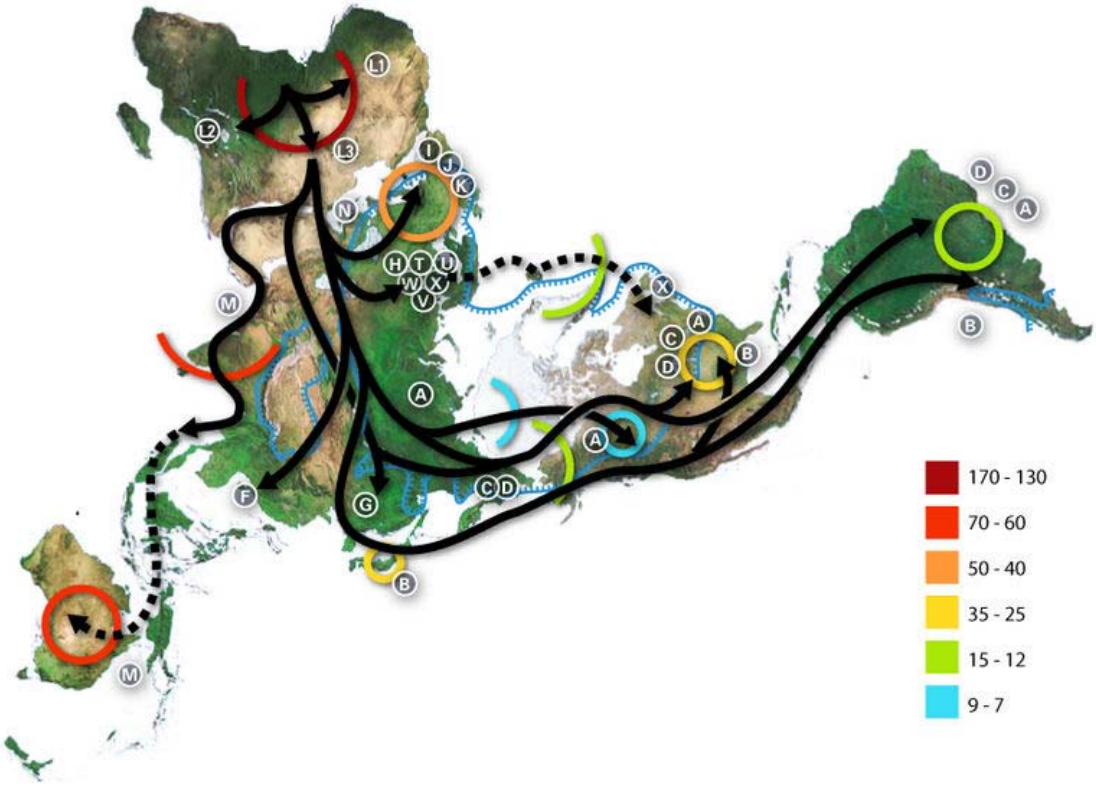
Few haplotypes will exactly match the modal values for Haplogroup G. One can consult an allele frequency table to determine the likelihood of remaining in Haplogroup G based on the variations observed.

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.

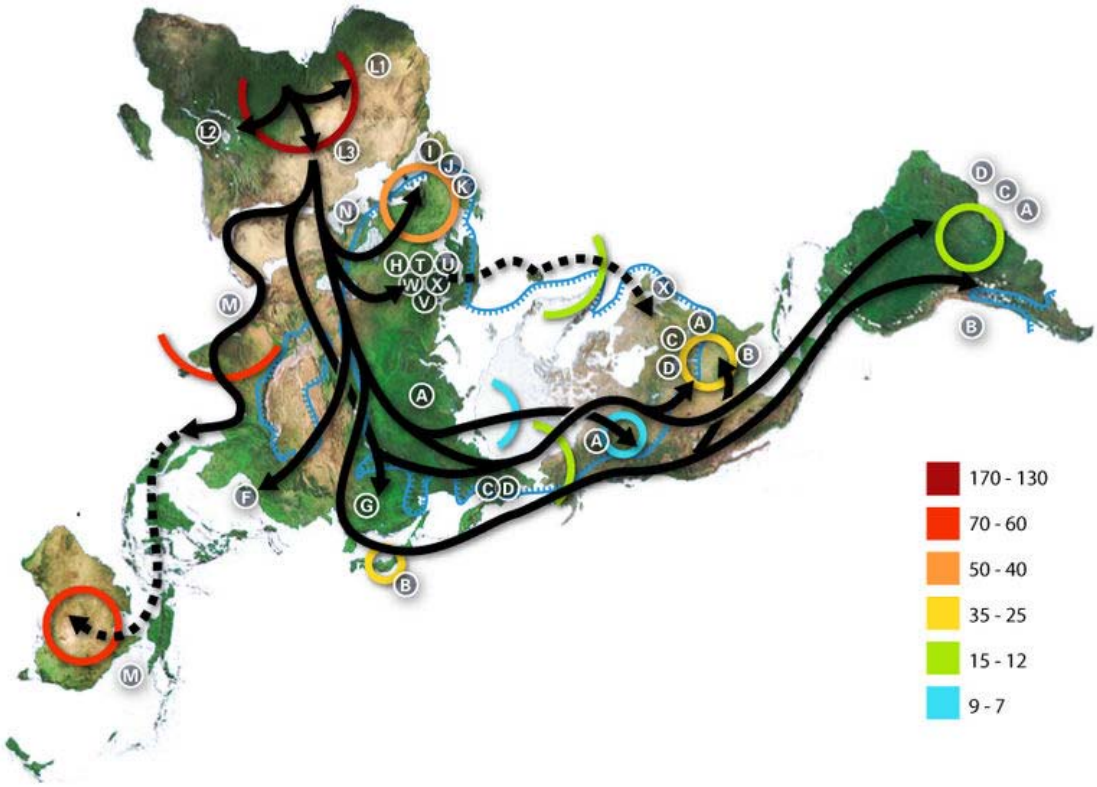
The testing of full mtDNA sequences is still somewhat controversial as it may reveal medical information. The field of linkage disequilibrium, unequal association of genetic disorders with a certain mitochondrial lineage, is in its infancy, but those mitochondrial mutations that have been linked are 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

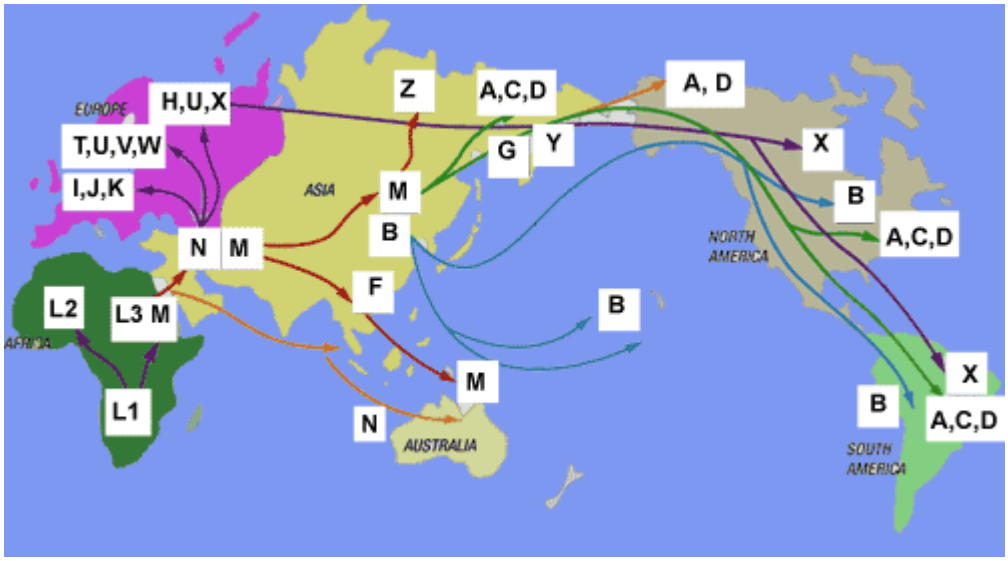
Some genealogy software programs now allow recording DNA marker test results, allowing for tracking of both Y-chromosome and mtDNA tests, and recording results for relatives. DNA-family tree wall charts are available.

Chapter- 3

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- 4

Human Y-chromosome DNA Haplogroup

In human genetics, a **Human Y-chromosome DNA haplogroup** is a haplogroup defined by differences in the non-recombining portions of DNA from the Y chromosome (called Y-DNA).

The Y chromosome consortium has established a system of defining Y-DNA haplogroups by letters A through to T, with further subdivisions using numbers and lower case letters.

Y-chromosomal Adam is the name given by researchers to a theoretical male who is the most recent common patrilineal (male-lineage) ancestor of all living humans. Estimations of the date of this common ancestor have varied significantly in different studies.

Major Y-DNA haplogroups

Major Y-chromosome haplogroups include:

Groups A and B

Haplogroups A and B are only found in sub-Saharan Africa (and in populations extracted from there in modern times, primarily via the Atlantic slave trade and Arab slave trade). The first to branch off was A, with defining mutation M91. All other haplogroups are summarized as BT (also referred to as YxA).

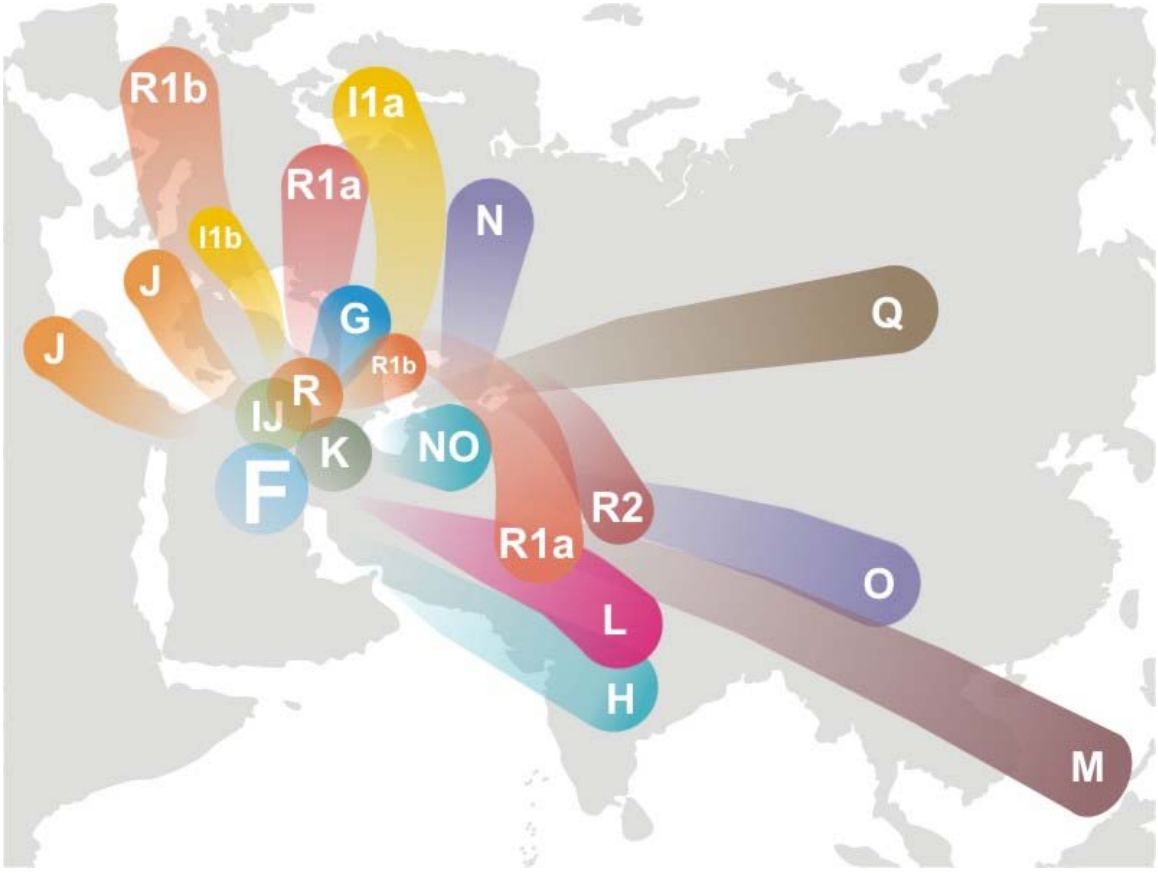
- Haplogroup A (M91) *Found in Africa, especially the Khoisan, Ethiopians (especially Beta Israel) and Nilotes*
- BT (M42, M94, M139, M299) ca. 55 ka BP
    - Haplogroup B (M60) *Found in Africa, especially the Pygmies and Hadzabe*
    - CT

Groups with mutation M168 (CT)

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

- Haplogroup CF (P143) *Found outside of Africa, throughout Eurasia, Oceania, and the Americas*
    - Haplogroup C (M130, M216) *Found in Asia, Oceania, and North America*
        - Haplogroup C1 (M8, M105, M131) *Found in Japan*
        - Haplogroup C2 (M38) *Found in Indonesia, New Guinea, Melanesia, Micronesia, and Polynesia*
        - Haplogroup C3 (M217, P44) *Found throughout Eurasia and North America, but especially among Mongols, Kazakhs, Tungusic peoples, Paleosiberians, and Na-Dené-speaking peoples*
        - Haplogroup C4 (M347) *Found among the indigenous peoples in Australia*
        - Haplogroup C5 (M356) *Found in the Indian subcontinent, the Arabian Peninsula, and northern China*
    - Haplogroup F (M89, M213) *Found in Southern India, Sri Lanka, Yunnan, Korea*
        - GT
- Haplogroup DE (M1, M145, M203) ca. 65 ka
    - Haplogroup D (M174) *Found in Japan, China (especially Tibet), the Andaman Islands*
        - Haplogroup D1 (M15)
        - Haplogroup D2 (M55, M57, M64.1, M179, P12, P37.1, P41.1 (M359.1), 12f2.2)
        - Haplogroup D3 (P47)
    - Haplogroup E (M40, M96) *Found primarily in Africa*
        - Haplogroup E1 (P147)
            - Haplogroup E1a (M33, M132) **formerly E1**
            - Haplogroup E1b (P177)
                - Haplogroup E1b1 (P2, DYS391p); **formerly E3**
                    - Haplogroup E1b1a (M2) *Found in Africa*; **formerly E3a**
                    - Haplogroup E1b1b (M215) *Found in East Africa, North Africa, the Middle East, and Europe (especially in areas near the Mediterranean)*; **formerly E3b**
        - Haplogroup E2 (M75)

Groups descended from Haplogroup F (G, H & IJK)



The diversion of Haplogroup F and its descendants.

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

- Haplogroup G (M201) ca. 21 ka *Found in many ethnic groups in Eurasia; most common in the Caucasus, Iran, Anatolia and other eastern Mediterranean coastal areas. Found in almost all European countries, but most common in Gagauzia, southeastern Romania, Greece, Italy, Spain, Portugal, Tyrol, and Bohemia with highest concentrations on some Mediterranean islands; uncommon in Northern Europe* Found in small numbers in northwestern China and India, Pakistan, Sri Lanka, Malaysia, and North Africa
    - Haplogroup G1
    - Haplogroup G2

        - Haplogroup G2a
            - Haplogroup G2a1
            - Haplogroup G2a2
            - Haplogroup G2a3
                - Haplogroup G2a3a
                - Haplogroup G2a3b
                    - Haplogroup G2a3b1
        - Haplogroup G2b
        - Haplogroup G2c (formerly Haplogroup G5)
            - Haplogroup G2c1
- Haplogroup H (M69) *Found mainly in South Asia*
    - Haplogroup H1
    - Haplogroup H2
- Haplogroup IJK
    - Haplogroup IJ (P123, P124, P125, P126, P127, P129, P130) ca. 45 ka
        - Haplogroup I (M170, M258) *Found in Europe and parts of the Near East*
            - Haplogroup I1 (M253) *Found mainly in northern Europe*
            - Haplogroup I2 (P215) *Found mainly in southeast Europe and Sardinia* save for I2B1 (m223) which is primarily found in Western, Central, and Northern Europe.
        - Haplogroup J (M304, S6, S34, S35)
            - Haplogroup J* (Rare outside of Socotra)
            - Haplogroup J1 *Associated with Northeast Caucasian peoples in Dagestan and Semitic peoples in Mesopotamia, the Levant, the Arabian Peninsula, Ethiopia, and North Africa*
            - Haplogroup J2 (M172) *Found mainly in Mesopotamia, Anatolia, Levant, Greece, the Balkans, Italy and the Caucasus*
    - Haplogroup K (M9) *Spread all over Eurasia, Oceania and Americas*
        - LT

Groups descended from Haplogroup K (LT)

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

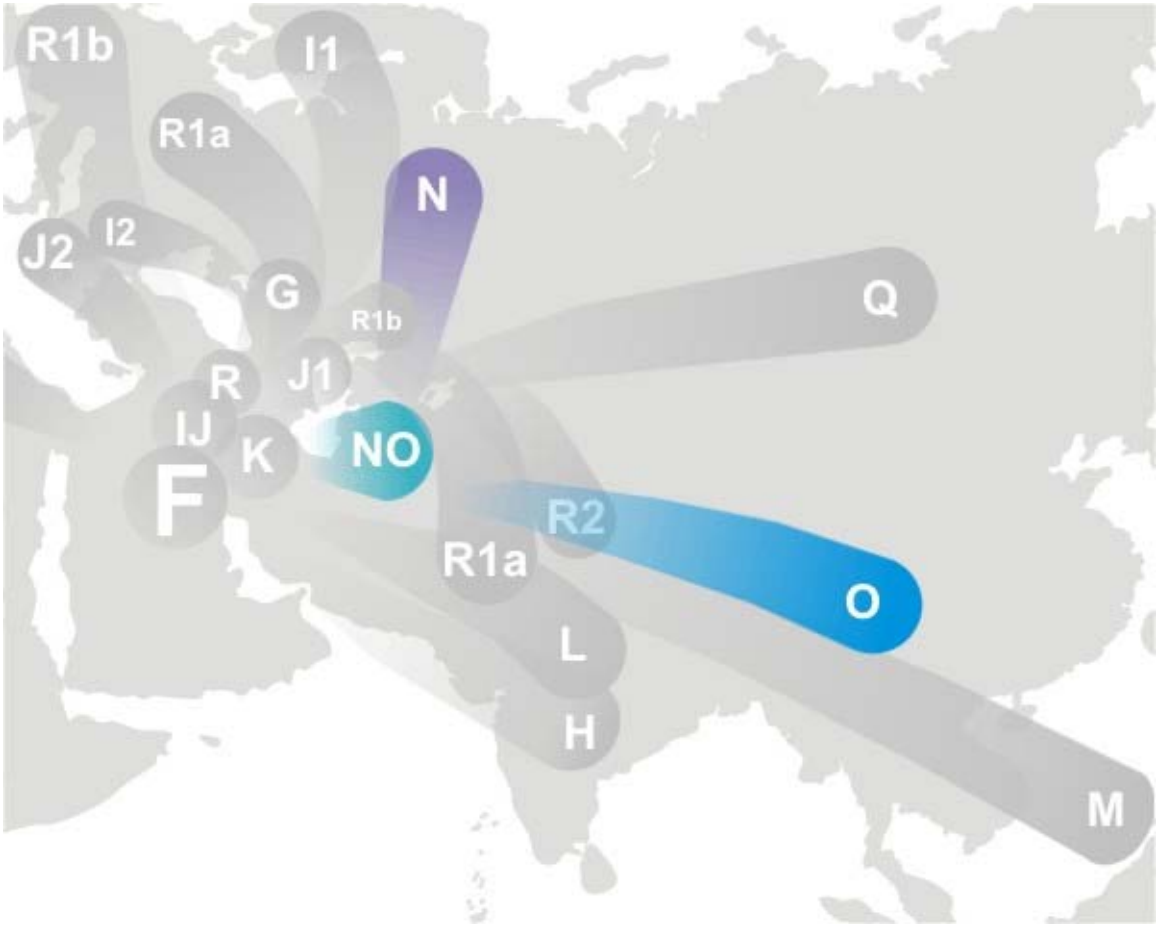
- Haplogroup K* *Found in Melanesia and Australia*
- Haplogroup K1 (formerly Haplogroup K3) *Found in Indian subcontinent*

- Haplogroup K2 (formerly Haplogroup K4)
- Haplogroup K3 (formerly Haplogroup K6) *Found in Melanesia and Polynesia*
- Haplogroup K4 *Found in Bali*

- Haplogroup L (M20) *Found in South Asia, Central Asia, Southwest Asia, the Mediterranean*
- Haplogroup MNOPS (rs2033003/M526)
    - Haplogroup M (P256) *Found in New Guinea and Melanesia*
    - Haplogroup NO (M214) 35-40 kya
        - Haplogroup NO* (minimal distribution)
        - Haplogroup N (M231) *Found in northernmost Eurasia, especially among the Uralic peoples*
        - Haplogroup O (M175) *Found in East Asia, Southeast Asia, the South Pacific*
    - Haplogroup P (M45)
        - Haplogroup P* (minimal distribution)
        - Haplogroup Q (MEH2, M242, P36) *Found in Siberia and the Americas*
        - Haplogroup R (M207, M306) *Found in Europe, West Asia, Central Asia, and South Asia*
    - Haplogroup S (M230) (formerly known as **Haplogroup K5**) *Found in the highlands of New Guinea*

- Haplogroup T (formerly known as **Haplogroup K2**) (M184, M70, M193, M272) *Found in Africa (mainly Afro-Asiatic-speaking peoples), the Middle East, the Mediterranean, South Asia. Found in a significant minority of Sciaccensi, Somalis, Eivissencs, Stilfser, Ethiopians, Fulbe, Egyptians, and Omanis; also found at low frequency throughout the Mediterranean and parts of India*

Groups descended from Haplogroup NO (M214)



The diversion of Haplogroup NO and its descendants.

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

- Haplogroup NO (M214) 35-40 ka (minimal distribution)
  - Haplogroup N (M231) *Found in northernmost Eurasia, especially among the Uralic peoples*
    - Haplogroup N1 (LLY22g)
  - Haplogroup O (M175) *Found in East Asia, Southeast Asia, the South Pacific*
    - Haplogroup O1 (MSY2.2) *Found in eastern and southern China, Taiwan, and Southeast Asia, especially among Austronesian and Kradai peoples*

        - Haplogroup O2 (P31, M268)
            - Haplogroup O2a (M95) *Found in Japan, southern China, Southeast Asia, and the Indian subcontinent, especially among Austro-Asiatic peoples, Kradai peoples, Malays, and Indonesians*
            - Haplogroup O2b (SRY465, M176) *Found in Japan, Korea, Manchuria, and Southeast Asia*
        - Haplogroup O3 (M122) *Found throughout East Asia, Southeast Asia, and Austronesia including Polynesia*

Groups descended from Haplogroup P (M45)

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

Haplogroup Q

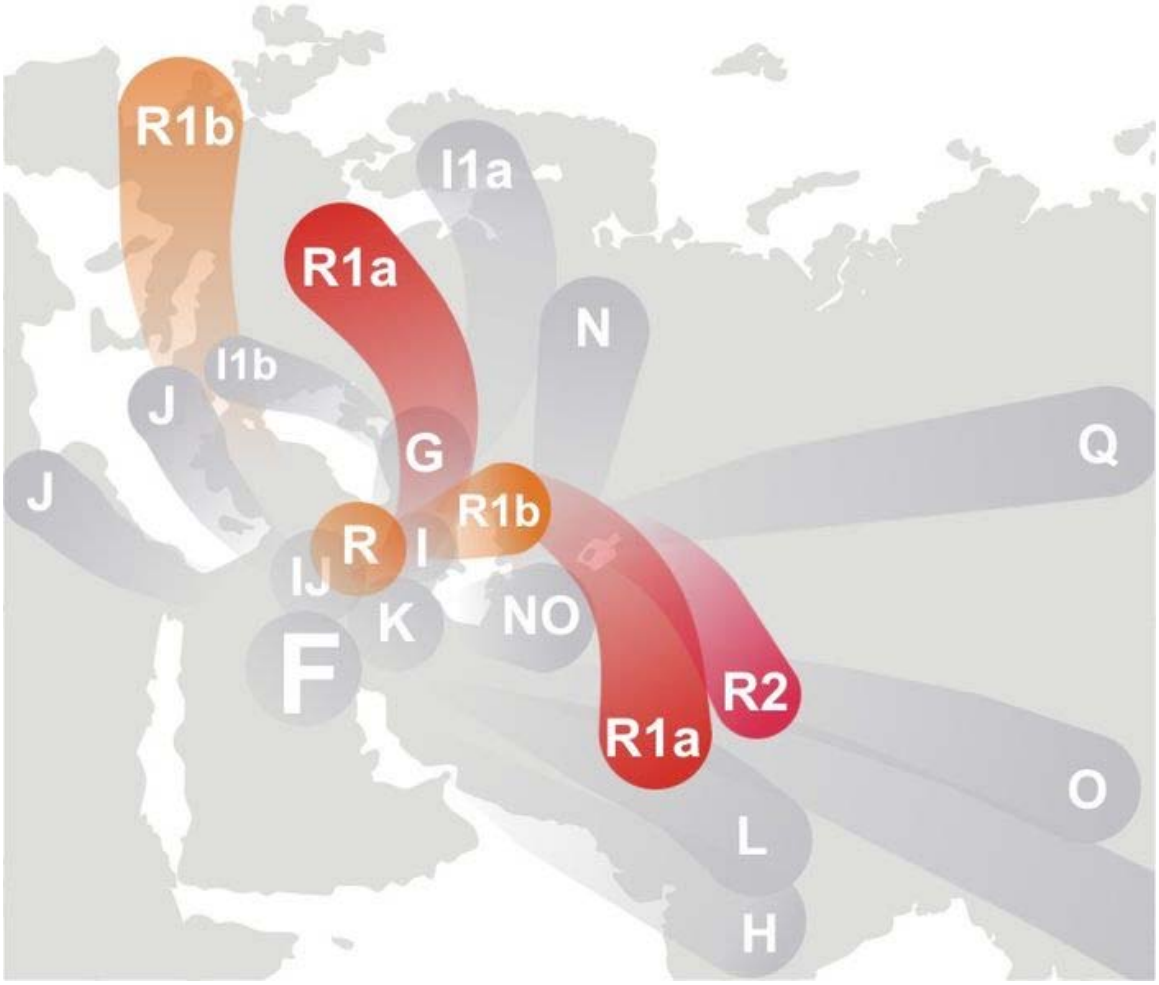
Q is defined by the SNP M242. It is believed to have arisen in Central Asia approximately 35-40 000 years ago. The subclades of Haplogroup Q with their defining mutation(s), according to the 2008 ISOGG tree are provided below. ss4 bp, rs41352448, is not represented in the ISOGG 2008 tree because it is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

The 2008 ISOGG tree

- Q (M242)
    - Q*
    - Q1 (P36.2)
        - Q1*
        - Q1a (MEH2)
            - Q1a*
            - Q1a1 (M120, M265/N14) *Found with low frequency among Dungans, Han Chinese, Hazaras, Japanese, Koreans, and Tibetans*
            - Q1a2 (M25, M143) *Found at low to moderate frequency among some populations of Southwest Asia, Central Asia, and Siberia*
            - Q1a3 (M346)
                - Q1a3* *Found at low frequency in Pakistan, India, and Tibet*
                - Q1a3a (M3) *Typical of indigenous peoples of the Americas*
                    - Q1a3a*
                    - Q1a3a1 (M19) *Found among some indigenous peoples of South America, such as the Ticuna and the Wayuu*

        - Q1a3a2 (M194)
        - Q1a3a3 (M199, P106, P292)
    - Q1a4 (P48)
    - Q1a5 (P89)
    - Q1a6 (M323) *Found in a significant minority of Yemeni Jews*
- Q1b (M378) *Found at low frequency among samples of Hazara and Sindhis*

Haplogroup R



The diversion of Haplogroup R and its descendants.

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

most commonly found among modern European populations, especially those of Western Europe.

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

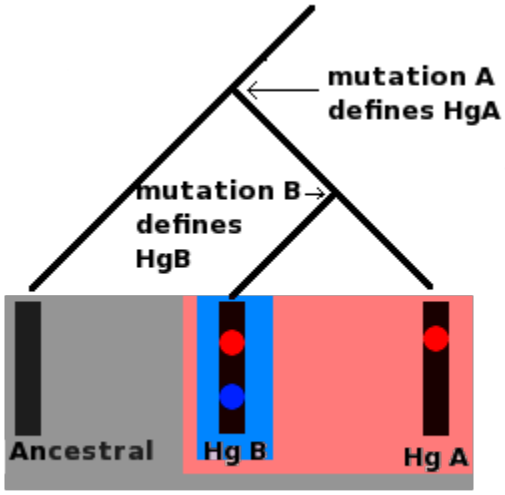
Chapter- 5

Haplogroup

In the study of molecular evolution, a **haplogroup** (from the Greek: ἁπλοῦς, *haploûs*, "onefold, single, simple") is a group of similar haplotypes that share a common ancestor with a single nucleotide polymorphism (SNP) mutation. Because a haplogroup consists of similar haplotypes, this is what makes it possible to predict a haplogroup from haplotypes. An SNP test confirms a haplogroup. Haplogroups are assigned letters of the alphabet, and refinements consist of additional number and letter combinations, for example R1b1. Y-chromosome and mitochondrial DNA haplogroups have different haplogroup designations. Haplogroups pertain to deep ancestral origins dating back thousands of years.

In human genetics, the haplogroups most commonly studied are Y-chromosome (Y-DNA) haplogroups and mitochondrial DNA (mtDNA) haplogroups, both of which can be used to define genetic populations. Y-DNA is passed solely along the patrilineal line, from father to son, while mtDNA is passed down the matrilineal line, from mother to offspring of both sexes. Neither recombines, and thus Y-DNA and mtDNA change only by chance mutation at each generation with no intermixture between parents' genetic material.

Haplogroup formation



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 eucaryotic cells, such as those of humans. Their primary purpose is to provide energy to the cell. Mitochondria are thought to be reduced descendants of symbiotic bacteria that were once free living. One indication that mitochondria were once free living is that each contains a circular DNA, called mitochondrial DNA (mtDNA), whose structure is more similar to bacteria than eukaryotic organisms. The overwhelming majority of a human's DNA is contained in the chromosomes in the nucleus of the cell, but mtDNA is an exception.

An individual inherits their cytoplasm and the organelles it contains exclusively from their mother, as these are derived from the ovum (egg cell), sperm only carry chromosomal DNA due to the necessity of maintaining motility. When a mutation arises in mtDNA molecule, the mutation is therefore passed in a direct female line of descent. These mutations are derived from copying mistakes, when the DNA is copied it is possible that a single mistake occurs in the DNA sequence, these single mistakes are called single nucleotide polymorphisms (SNPs).

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) arise in the Y chromosome, they are passed on directly from father to son in a direct male line of descent. The Y chromosome and mtDNA therefore share specific properties.

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 chromosomes and mtDNA display is that mutations can accrue along a certain segment of both molecules and these mutations remain fixed in place on the 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 person 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 mutation defines a set of specific Y chromosomes called a haplogroup.

All men carrying mutation A form a single haplogroup, all men carrying mutation B are part of this haplogroup, but mutation B also defines 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 are grouped into lineages and haplogroups; these are often presented as tree like diagrams.

Haplogroup population genetics

It is usually assumed that there is little natural selection for or against a particular haplotype mutation which has survived to the present day, so apart from mutation rates (which may vary from one marker to another) the main driver of population genetics affecting the proportions of haplotypes in a population is genetic drift — random fluctuation caused by the sampling randomness of which members of the population happen to pass their DNA on to members of the next generation of the appropriate sex.

This causes the prevalence of a particular marker in a population to continue to fluctuate, until it either hits 100%, or falls out of the population entirely. In a large population with efficient mixing the rate of genetic drift for common alleles is very low; however, in a

very small interbreeding population the proportions can change much more quickly. The marked geographical variations and concentrations of particular haplotypes and groups of haplotypes therefore witness the distinctive effects of repeated population bottlenecks or founder events followed by population separations and increases.

The lineages which can be traced back from the present will not reflect the full genetic variation of the older population: genetic drift means that some of the variants will have died out. The cost of full Y-DNA and mtDNA sequence tests has limited the availability of data; however, their cost has dropped dramatically in the last decade. Haplotype coalescence times and current geographical prevalences both carry considerable error uncertainties. This is especially troublesome for coalescence times, because most population geneticists still continue (albeit decresing a little bit) to use the "Zhivotovski method", which is heavily criticised by DNA-genealogists for its' falsehood.

Groups without mutation M168

- Haplogroup A (M91) (Africa, especially the Khoisan, Ethiopians, and Nilotes)
- Haplogroup B (M60) (Africa, especially the Pygmies and Hadzabe)

Groups with mutation M168

(mutation M168 occurred ~50,000 bp)

- Haplogroup C (M130) (Oceania, North/Central/East Asia, North America and significant presence in India)
- Haplogroup F (M89) Oceania, Europe, Asia, North- and South- America
- YAP+ haplogroups
    - Haplogroup DE (M1, M145, M203)
        - Haplogroup D (M174) (Tibet, Japan, the Andaman Islands)
        - Haplogroup E (M96)
            - Haplogroup E1b1a (M2) West and Central Africa and surrounding regions; formerly known as E3a
            - Haplogroup E1b1b (M35) East Africa, North Africa, the Middle East, the Mediterranean, the Balkans; formerly known as E3b

Groups with mutation M89

(mutation M89 occurred ~45,000 bp)

- Haplogroup F (P14, M213) (southern India, Sri Lanka, China, Korea)
- Haplogroup G (M201) (present among many ethnic groups in Eurasia, usually at low frequency; most common in the Caucasus, the Iranian plateau, and Anatolia;

in Europe mainly in Greece, Italy, Iberia, the Tyrol, Bohemia; extremely rare in Northern Europe)

- Haplogroup H (M69) (India, Sri Lanka, Nepal, and at low frequency in Pakistan, Iran, Central Asia, and Arabia)

- Haplogroup IJK (L15, L16)

Groups with mutations L15 & L16

- Haplogroup IJK (L15, L16)
    - Haplogroup IJ (S2, S22)
        - Haplogroup I (M170, P19, M258) (widespread in Europe, found infrequently in parts of the Middle East, and virtually absent elsewhere)
            - Haplogroup I1 (M253, M307, P30, P40) (Northern Europe)
            - Haplogroup I2 (S31) (Central and Southeast Europe, Sardinia)
        - Haplogroup J (M304) (the Middle East, Turkey, Caucasus, Italy, Greece, the Balkans, North and Northeast Africa)
            - Haplogroup J* (Mainly found in Socotra, with a few observations in Pakistan, Oman, Greece, Czechia, and among Turkic peoples)
            - Haplogroup J1 (M267) (Mostly associated with Semitic peoples in the Middle East, Ethiopia, and North Africa, and with Northeast Caucasian peoples in Dagestan; J1 with DYS388=13 is associated with eastern Anatolia)
            - Haplogroup J2 (M172) (Mainly found in West Asia, Central Asia, South Asia, Southern Europe, and North Africa)
    - Haplogroup K (M9, P128, P131, P132)

Groups with mutation M9

(mutation M9 occurred ~40,000 bp)

- Haplogroup K
    - Haplogroup L (M11, M20, M22, M61, M185, M295) (South Asia, Central Asia, Southwestern Asia, the Mediterranean)
    - Haplogroup MNOPS (rs2033003/M526)
    - Haplogroup T (M70, M184/USP9Y+3178, M193, M272)

Groups with mutation M526

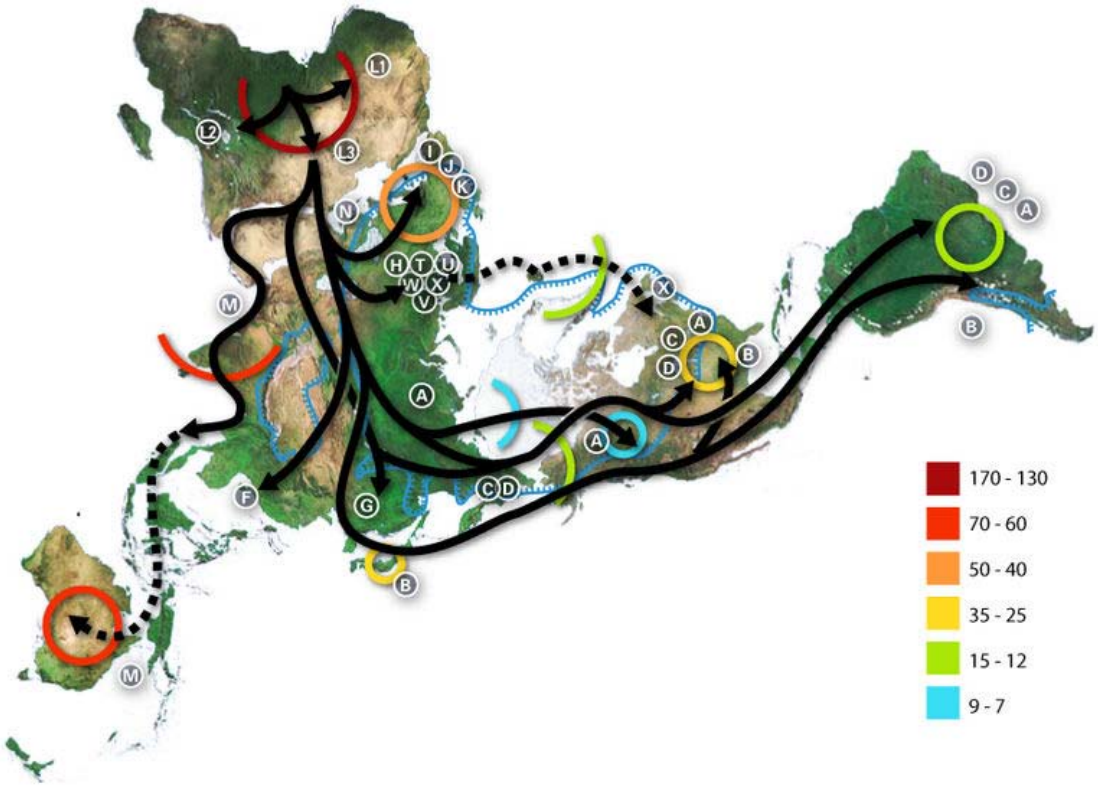
- Haplogroup MNOPS (rs2033003/M526)
    - Haplogroup M (P256) (New Guinea, Melanesia, eastern Indonesia)

- Haplogroup NO (M214)
  - Haplogroup N (M231) (northernmost Eurasia, especially among the Uralic peoples)
  - Haplogroup O (M175) (East Asia, Southeast Asia, the South Pacific, South Asia, Central Asia)
    - Haplogroup O1 (MSY2.2)
    - Haplogroup O2 (P31, M268)
    - Haplogroup O3 (M122)
- Haplogroup P (M45, 92R7, M74/N12) (M45 occurred ~35,000 bp)
  - Haplogroup Q (MEH2, M242, P36) (Occurred ~15,000-20,000 years ago. Found in Asia and the Americas)
    - Haplogroup Q1a3a1 (M3) (North America, Central America, and South America)
  - Haplogroup R (M207)
    - Haplogroup R1 (M173)
      - Haplogroup R1a (M17) (Central Asia, South Asia, and Central, Northern, and Eastern Europe)
      - Haplogroup R1b (M343) (Europe, Caucasus, Central Asia, South Asia, North Africa, Central Africa)
    - Haplogroup R2 (M124) (South Asia, Caucasus, Central Asia)
- Haplogroup S (M230, P202, P204) (New Guinea, Melanesia, eastern Indonesia)

Groups with mutation M70

- Haplogroup T - (North Africa, Horn of Africa, Southwest Asia, the Mediterranean, South Asia); formerly known as Haplogroup K2

Defining populations



Map of human haplotype migration, according to mitochondrial DNA.

Haplogroups can be used to define genetic populations and are often geographically oriented. For example, the following are common divisions for mtDNA haplogroups:

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

The mitochondrial haplogroups are divided into 3 main groups, which are designated by the 3 sequential letters L, M, N. Humanity first split within the L group between L0 and L1-6. L1-6 gave rise to other L groups, one of which, L3, split into the M and N group. The M group comprises the first wave of human migration out of Africa, following an eastward route along southern coastal areas.

Descendent populations belonging to haplogroup M are found throughout East Africa, Asia, the Americas, and Melanesia, though almost none have been found in Europe. The

N group may represent another migration out of Africa, heading northward instead of eastward. Shortly after the migration, the large R group split off from the N.

Haplogroup R consists of two subgroups defined on the basis of their geographical distributions, one found in southeastern Asia and Oceania and the other containing almost all of the modern European populations. Haplogroup N(xR), i.e. mtDNA that belongs to the N group but not to its R subgroup, is typical of Australian aboriginal populations, while also being present at low frequencies among many populations of Eurasia and the Americas.

The L type consists of nearly all Africans.

The M type consists of:

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

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

D- Some Amerindians, many Siberians and northern East Asians

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

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

Q- Melanesian, Polynesian, New Guinean populations

The N type consists of:

A- Found in some Amerindians, Japanese, and Koreans

I- 10% frequency in Northern, Eastern Europe

S- Some Australian aborigines

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

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

Y- Most Nivkhs and many Ainus; 1% in Southern Siberia

R- Large group found within the N type.Populations contained therein can be divided geographically into West Eurasia and East Eurasia. Almost all European populations and

a large number of Middle-Eastern population today are contained within this branch. A smaller percentage is contained in other N type groups). Below are **subclades of R**:

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

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

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

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

U- High frequency in Scandinavia, Baltic countries, Mediterranean

Overlap between y-haplogroups and mt-haplogroups

The ranges of specific y-haplogroups and specific mt-haplogroups overlap, indicating populations that have a specific combination of a y-haplogroup and an mt-haplogroup. Y mutations and mt mutations do not necessarily occur at a similar time, and differential rates of sexual selection between the two genders combined with founder effect and genetic drift can alter the haplogroup composition of a population, so the overlaps are only rough.

The very rough overlaps between Y-DNA haplogroups and mtDNA haplogroups are as follows:

Y-DNA haplogroup(s)	mtDNA haplogroup(s)	Geographical area and/or peoples
A	L0	Southern Africa, Khoisan
B	L1	Middle Africa
E	L2, L3	Africa
O, N, C3	CZ/C/Z, D, G (M types); A (N type); B, F (R types)	East Asia, Siberia
K, M (M9-positive, M45-negative)	B, P (R types); N; Q (M type) as well as various Oceanian-specific M subclades	Oceania
R, I, T, J, E (V13, M81, and M123 types)	R0, HV/H/V, JT/J/T, U/K (R types)	Europe, West Asia, North Africa, Horn of Africa
Q, C3	A, X (N types); C, D (M types)	Easternmost Siberia, the Americas

Chapter- 6

Haplotype

A **haplotype** (from the Greek: ἁπλοῦς, *haploûs*, "onefold, single, simple") in genetics is a combination of alleles (DNA sequences) at different places (loci) on the chromosome that are transmitted together. A haplotype may be one locus, several loci, or an entire chromosome depending on the number of recombination events that have occurred between a given set of loci.

In a second meaning, haplotype is a set of single-nucleotide polymorphisms (SNPs) on a single chromosome of a chromosome pair that are statistically associated. It is thought that these associations, and the identification of a few alleles of a haplotype block, can unambiguously identify all other polymorphic sites in its region. Such information is very valuable for investigating the genetics behind common diseases, and has been investigated in the human species by the International HapMap Project.

Many genetic testing companies use the term 'haplotype' to refer to an individual collection of short tandem repeat (STR) allele mutations within a genetic segment, while using the term 'haplogroup' to refer to the SNP/unique-event polymorphism (UEP) mutations which represents the clade to which a collection of potential haplotypes belong.

Haplotype resolution

An organism's genotype may not uniquely define its haplotype. For example, consider a diploid organism and two bi-allelic loci on the same chromosome such as single-nucleotide polymorphisms (SNPs). The first locus has alleles *A* and *T* with three possible genotypes *AA*, *AT*, and *TT*, the second locus having *G* and *C*, again giving three possible genotypes *GG*, *GC*, and *CC*. For a given individual, there are therefore nine possible configurations for the genotypes at these two loci, as shown in the Punnett square below, which shows the possible genotypes that an individual may carry and the corresponding haplotypes that these resolve to. For individuals that are homozygous at one or both loci,

it is clear what the haplotypes are; it is only when an individual is heterozygous at both loci that the gametic phase is ambiguous.

	AA	**AT**	**TT**
GG	AG AG	AG TG	TG TG
GC	AG AC	AG TC or AC TG	TG TC
CC	AC AC	AC TC	TC TC

The only unequivocal method of resolving phase ambiguity is by sequencing. However, it is possible to estimate the probability of a particular haplotype when phase is ambiguous using a sample of individuals.

Given the genotypes for a number of individuals, the haplotypes can be inferred by haplotype resolution or haplotype phasing techniques. These methods work by applying the observation that certain haplotypes are common in certain genomic regions. Therefore, given a set of possible haplotype resolutions, these methods choose those that use fewer different haplotypes overall. The specifics of these methods vary - some are based on combinatorial approaches (e.g., parsimony), whereas others use likelihood functions based on different models and assumptions such as the Hardy-Weinberg principle, the coalescent theory model, or perfect phylogeny. These models are combined with optimization algorithms such as expectation-maximization algorithm (EM), Markov chain Monte Carlo (MCMC), or hidden Markov models (HMM).

Y-DNA haplotypes from genealogical DNA tests

Unlike other chromosomes, Y chromosomes do not come in pairs. Every human male has only one copy of that chromosome. This means that there is no lottery as to which copy to inherit, and also (for most of the chromosome) no shuffling between copies by recombination; so, unlike autosomal haplotypes, there is therefore effectively no randomisation of the Y-chromosome haplotype between generations, and a human male should largely share the same Y chromosome as his father, give or take a few mutations.

In particular, the Y-DNA that is the numbered results of a Y-DNA genealogical DNA test should match, barring mutations. Within genealogical and popular discussion, this is sometimes referred to as the "DNA signature" of a particular male human, or of his paternal bloodline.

UEP results (SNP results)

Unique-event polymorphisms (UEPs) like SNPs represent haplogroups. STRs represent haplotypes. The results that make up the full Y-DNA haplotype from the Y chromosome DNA test can be divided into two parts: the results for UEPs, sometimes loosely called the SNP results as most UEPs are single-nucleotide polymorphisms, and the results for

microsatellite short tandem repeat sequences (Y-STRs), often designated by DYS numbers.

The UEP results reflect the inheritance of events it is believed can be assumed to have happened only once in all human history. These can be used to directly identify the individual's Y-DNA haplogroup, his place on the broad family tree of the whole of humanity. Different Y-DNA haplogroups identify genetic populations which are often intricately geographically oriented, reflecting the migrations of current individuals' direct patrilineal ancestors tens of thousands of years ago.

Y-STR haplotypes

The other possible part of the genetic results is the **Y-STR haplotype**, the set of results from the Y-STR markers tested.

Unlike the UEPs, the Y-STRs mutate much more easily, which gives them much more resolution to distinguish recent genealogy. But it also means that, rather than the population of descendants of a genetic event all sharing the *same* result, the Y-STR haplotypes are likely to have spread apart, to form a *cluster* of more or less similar results. Typically, this cluster will have a definite most probable center, the **modal haplotype** (presumably close to the haplotype of the original founding event), and also a **haplotype diversity** — the degree to which it has become spread out. The further in the past the defining event occurred, and the more that subsequent population growth occurred early, the greater the haplotype diversity for a particular number of descendants will be. On the other hand, if the haplotype diversity is smaller for a particular number of descendants, this may indicate a more recent common ancestor, or that a population expansion has occurred more recently.

It is important to note that, unlike for UEPs, there is no guarantee that two individuals with a similar Y-STR haplotype will necessarily share a similar ancestry. There is no uniqueness about Y-STR events. Instead, the clusters of Y-STR haplotype results inheriting from different events and different histories all tend to overlap.

Thus, although sometimes a Y-STR haplotype may be directly indicative of a particular Y-DNA haplogroup, it is in most cases a long time since the haplogroups' defining events, so typically the cluster of Y-STR haplotype results associated with descendents of that event has become rather broad, and will tend to significantly overlap the (similarly broad) clusters of Y-STR haplotypes associated with other haplogroups, making it impossible to predict with absolute certainty to which Y-DNA haplogroup a Y-STR haplotype would point. All that can be done from the Y-STRs, if the UEPs are not actually tested, is to predict probabilities for haplogroup ancestry (as this online program does), but not certainties.

A similar scenario exists for surnames. A cluster of similar Y-STR haplotypes may indicate a shared common ancestor, with an identifiable modal haplotype, but only if the cluster is sufficiently distinct from what may have arisen by chance from different

individuals historically having adopted the same name independently. This may require the typing of quite an extensive haplotype to establish, which has fuelled DNA testing companies to offer ever-larger sets of markers - 24 then 37 then 67, and perhaps soon even more.

Plausibly establishing relatedness between different surnames data-mined from a database is significantly harder, because now it must be established not that a *randomly-selected* member of the population is unlikely to have such a close match by accident, but rather that the *very nearest* member of the population in question, chosen purposely from the population for that very reason, would even under those circumstances be unlikely to match by accident. This is for the foreseeable future likely to be impossible, except in special cases where there is further information to drastically limit the size of that population of candidates under consideration.

Chapter- 7

Most Recent Common Ancestor

In genetics, the **most recent common ancestor** (**MRCA**) of any set of organisms is the most recent individual from which all organisms in the group are directly descended. The term is often applied to human genealogy.

The MRCA of a set of individuals can sometimes be determined by referring to an established pedigree. However, in general, it is impossible to identify the specific MRCA of a set of individuals, but an estimate of the time at which the MRCA lived can often be given; such estimates can be given based on DNA test results and established mutation rates, or by reference to a non-genetic genealogical model. This time estimate is referred to as **TMRCA** in scientific papers.

The term *MRCA* is usually used to describe a common ancestor of individuals within a species. It can also be used to describe a common ancestor between species. To avoid confusion, **last common ancestor** (**LCA**) or the equivalent term **concestor** is sometimes used in place of MRCA when discussing ancestry between species.

The term *MRCA* may also be used to identify a common ancestor between a set of organisms via specific gene pathways. Mitochondrial Eve and Y-chromosomal Adam are examples of such MRCAs.

MRCA of all living humans

Tracing one person's lineage back in time forms a binary tree of parents, grandparents, great-grandparents and so on. However, the number of individuals in such an ancestor tree grows exponentially and will eventually become impossibly high. For example, an individual human alive today would, over 30 generations, going back to about the High Middle Ages, have 2^{30} or about 1.07 billion ancestors, more than the total world population at the time.

In reality, an ancestor tree is not a binary tree. Rather, pedigree collapse changes the binary tree to a directed acyclic graph.

Consider the formation, one generation at a time, of the ancestor graph of all currently living humans with no descendants. Start with living people with no descendants at the bottom of the graph. Adding the parents of all those individuals at the top of the graph will connect (half-) siblings via one or two common ancestors, their parent(s). Adding the next generation will connect all first cousins. As each of the following generations of ancestors is added to the top of the graph, the relationship between more and more people is mapped (second cousins, third cousins and so on). Eventually a generation is reached where one or more of the many top-level ancestors is an MRCA from whom it is possible to trace a path of direct descendants all the way down to every living person at the bottom generations of the graph.

The MRCA of everyone alive today could thus have co-existed with a large human population, most of whom either have no living descendants today or else are ancestors of a subset of people alive today. The existence of an MRCA does therefore not imply the existence of a population bottleneck or first couple.

It is incorrect to assume that the MRCA passed all of his or her genes (or indeed any single gene) down to every person alive today. Because of sexual reproduction, at every generation, an ancestor only passes half of his or her genes to each particular descendant in the next generation. Save for inbreeding, the percentage of genes inherited from the MRCA becomes smaller and smaller in individuals at every successive generation, sometimes even decreasing to zero (at which point the Ship of Theseus situation arises), as genes inherited from contemporaries of MRCA are interchanged via sexual reproduction.

Ways to find the MRCA

There are a number of ways to estimate the MRCA such as genetics, archaeology, mathematical models, computer simulations and History. DNA studies have a problem in telling us about the MRCA. As Chang notes, the MRCA will be much more recent than any MRCA that could ever be found in DNA studies, even if one were to study the ancestry of every single gene. The reason being that we are considering people who are simply ancestors, through any route, whether or not any of their genes actually survived the journey. As the human genome consists of roughly 2^{32} base pairs, the genetic contribution of a single ancestor may be flushed out of an individual's genome completely after 32 generations, or roughly 1,000 years.

Time estimates

Depending on the survival of isolated lineages without admixture from modern migrations and taking into account long-isolated peoples, such as historical tribes in central Africa, Australia, and remote islands in the South Pacific, the human MRCA was generally assumed to have lived in the Upper Paleolithic period.

According to Rohde and his colleagues, if we consider not just our all-female and all-male lines, but our ancestors along all parental lines, it turns out that everyone on earth may share a common ancestor who is remarkably recent. Rohde, Olson, and Chang (2004), using a non-genetic model, estimated that the MRCA of all living humans may have lived within historical times (3rd millennium BC to 1st millennium AD). The paper suggests, "No matter the languages we speak or the color of our skin, we share ancestors who planted rice on the banks of the Yangtze, who first domesticated horses on the steppes of the Ukraine, who hunted giant sloths in the forests of North and South America, and who labored to build the Great Pyramid of Khufu". Rohde (2005) refined the simulation with parameters from estimated historical human migrations and of population densities.

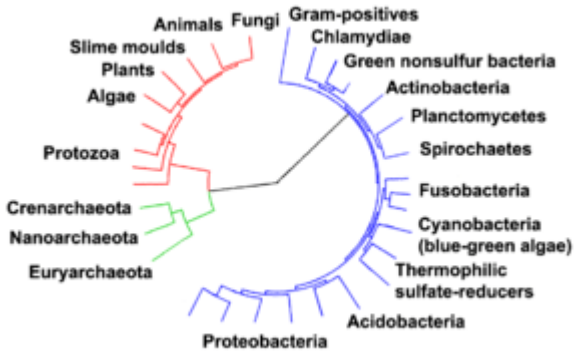
For conservative parameters, he pushes back the date for the MRCA to the 6th millennium BC (p. 20), but still concludes with a "surprisingly recent" estimate of a MRCA living in the second or first millennium BC (p. 27). An explanation of this result is that, while humanity's MRCA was indeed a Paleolithic individual up to early modern times, the European explorers of the 16th and 17th centuries would have fathered enough offspring so that some "mainland" ancestry by today pervades even remote habitats. Besides dating our most recent common ancestor, Rohde's team also calculates that in 5,400 BC everyone alive was either an ancestor of all of humanity, or of nobody alive today. The researchers call this the 'identical ancestors' point: the time before which all the family trees of people today are composed of exactly the same individuals.

However, the assumption that there are no isolated populations anymore is strongly questionable in view of the continuing existence of various uncontacted peoples, such as the Sentinelese, who are suspected to have been isolated completely (not only from the western world, but also from the Asian mainland in the case of the Sentinelese) possibly for many millennia.

Other models reported in Rohde, Olson, and Chang (2004) suggest that the MRCA of Western Europeans and people of Western European ancestry lived as recently as AD 1000. The same article provides surprisingly recent estimates for the identical ancestors point, the most recent time when each person then living was either an ancestor of all the persons alive today or an ancestor of none of them. The estimates for this are similarly uncertain, but date to considerably earlier than the MRCA, according to Rohde (2005) roughly to between 15,000 and 5,000 years ago.

A 2008 study found only six women contributed mitrochondrial DNA for 95% of all surviving Native Americans, and examination of mutation rates has been used to date waves of migration.

MRCA of different species



Evolutionary tree showing the divergence of modern species from the last universal ancestor in the center. The three domains are colored, with bacteria blue, archaea green, and eukaryotes red.

It is also possible to use the term MRCA to describe the common ancestor of two or more different species. In the past, the term MRCA was used interchangeably with **last common ancestor** (**LCA**) to denote both the common ancestor within a species and that between species. But MRCA is now more frequently used to describe common ancestors within a species. On the other hand, LCA now describes the common ancestor between two species.

The concept of the last common ancestor is described in Richard Dawkins' book, *The Ancestor's Tale*, in which he imagines a 'pilgrimage' backwards in time, during which we humans travel back through our own evolutionary history and as we do so are joined at each successive stage by all the other species of organism with which we share each respective common ancestor. Dawkins uses the word "**concestor**" (coined by Nicky Warren) as an alternative to LCA.

In *The Ancestor's Tale*, following the human evolutionary tree backwards, we first meet the concestor which we share with the species that are our closest relatives, the chimpanzee and bonobo. Dawkins estimates this to have occurred between 5 and 7 million years ago. Another way of looking at this is to say that our (approximately) 250,000-greats-grandparent was a creature from which all humans, chimpanzees and bonobos are directly descended. Further on in Dawkins' imaginary journey, we meet the concestor we share with the Gorilla, our next nearest relative, then the Orangutan, and so on, until we finally meet the concestor of all living organisms, known as the last universal ancestor.

A common mistake is to refer to a proposed *last* common ancestor as an *earliest* ancestor (as in the book *The Link: Uncovering Our Earliest Ancestor* by Colin Tudge, and the documentary *Uncovering Our Earliest Ancestor: The Link* screened on the History Channel (US) and BBC One (UK), both referring to the primate fossil dubbed Ida).

MRCA of a population identified by a single genetic marker

It is also possible to consider the ancestry of individual genes (coalescent theory), instead of a person (an organism) as a whole. Unlike organisms, a gene is passed down from a generation of organisms to the next generation either as perfect replicas of itself or as slightly mutated *descendant genes*. While organisms have ancestry graphs and progeny graphs via sexual reproduction, a gene has a single chain of ancestors and a tree of descendants. An organism produced by non-autogamous sexual reproduction has at least two ancestors (immediate parents), but a gene always has one single ancestor.

Given any gene in the body of a person, we can trace a single chain of *human ancestors* back in time, following the lineage of this one gene. Because a typical organism is built from tens of thousands of genes, there are numerous ways to trace the ancestry of organisms using this mechanism. But all these inheritance pathways share one common feature. If we start with all humans alive in 1995 and trace their ancestry by one particular gene (actually a locus), we find that the farther we move back in time, the smaller the number of ancestors becomes. The pool of ancestors continues to shrink until we find the *most recent common ancestor* (MRCA) of all humans who were alive in 1995, *via this particular gene pathway*.

Patrilineal and matrilineal MRCA

It is not possible to trace human ancestry via autosomal chromosomes. Although a chromosome contains genes that are passed down from parents to children via independent assortment from only one of the two parents, genetic recombination (chromosomal crossover) mixes genes from non-sister chromatids from both parents during meiosis, thus changing the nature of the chromosome. In addition, each parent will pass on only one of their autosomal chromosomes to their offspring.

However, the mitochondrial DNA (mtDNA) is nearly immune to sexual mixing. Mitochondrial DNA, therefore, can be used to trace matrilineal inheritance of a population of related individuals. Similarly Y-DNA is present as a single chromosome in the male individual and is passed on to sons and grandsons without recombination.

Time estimates or TMRCA

Both Y-DNA and mtDNA are thus used to trace ancestry. Populations are defined by the accumulation of mutations on the gene and special trees are created for the mutations and the order in which they occurred in a population. The tree is formed through the testing of a large number of individuals all over the world for the presence or lack of a certain set of mutations. Once this is done it is possible to determine how many mutations separate one population from another in the case of mtDNA. The number of mutations in turn allow scientists to determine the approximate time passed, or the TMRCA, since the populations separated. This estimate is based on the estimated mutation rate of the mtDNA in the regions tested. The TMRCA would be the time when both populations still shared the same set of mutations or belonged to the same haplogroup.

In the case of Y-DNA TMRCA is arrived at in a different way. Haplogroups are defined by single-nucleotide polymorphism in various regions of the Y-DNA. The TMRCA within the haplogroup is defined by the accumulation of mutations in STR sequences of the Y-Chromosome of that haplogroup only. Y-DNA network analysis of Y-STR haplotypes yielding a star cluster can be regarded as representing a population descended from a single ancestor. In this case the variability of the DYS sequence, also called the microsatellite variation, can be regarded as a measure of the time passed since the ancestor founded this particular population. Variability due to multiple founding individuals will not display as a star cluster and overall variability in the Y-DNA is thus due in part to variability in the original founding population. The descendants of Genghis Khan or one of his ancestors represents a famous star cluster than can be dated back to the time of Genghis Khan.

Mitochondrial Eve and the most recent common patrilineal ancestor of all living male humans, known as Y-chromosomal Adam, have been established by researchers using tests of the same kinds of DNA as for two individuals. Mitochondrial Eve is estimated to have lived about 140,000 years ago. Y-chromosomal Adam is estimated to have lived around 60,000 years ago. The MRCA of humans alive today would therefore need to have lived more recently than either.

TMRCA calculations are considered critical evidence when attempting to determine migration dates of various populations as they spread around the world. For example, if a mutation is deemed to have occurred 30,000 years ago, then this mutation should be found amongst all populations that diverged after this date. If archeological evidence indicates cultural spread and formation of regionally isolated populations then this must be reflected in the isolation of subsequent genetic mutations in this region. If genetic divergence and regional divergence coincide it can be concluded that the observed divergence is due to migration as evidenced by the archaeological record. However, if the date of genetic divergence occurs at a different time than the archaeological record, then scientists will have to look at alternate archaeological evidence to explain the genetic divergence. The issue is best illustrated in the debate surrounding the demic diffusion versus cultural diffusion during the European Neolithic.

Identical ancestors point

The MRCA had many contemporary companions of both sexes. Many of these contemporaries left direct descendants, but not all of them left an unbroken link of descendants all the way down to today's population. That is, some contemporaries of the MRCA are ancestors of no one in the current population. The rest of the contemporaries of the MRCA may claim ancestry over a subset of current population.

Because ancestors of MRCA are by definition also common ancestors, we can continue to find (less recent) common ancestors by pushing further back in time to the MRCA's ancestors. Eventually we reach a point in the past where all humans can be divided into two groups: those who left no descendants today and those who are common ancestors of all living humans today. This point in time is termed the *identical ancestors point*. Even

though each living person receives genes (in original or mutated forms) in dramatically different proportions from these ancestors from the identical ancestors point, from this point back, all living people share exactly the same set of ancestors, all the way to the very first single-celled organism.

Chapter- 8

Personal Genomics

Personal genomics is a branch of genomics where individual genomes are genotyped and analyzed using bioinformatics tools. It is also related to traditional population genetics. The genotyping stage can have many different experimental approaches including single nucleotide polymorphism (SNP) chips (typically 0.02% of the genome), or partial or full genome sequencing. Once the genotypes are known, there are many bioinformatics analysis tools that can compare individual genomes and find disease association of the genes and loci. The most important aspect of personal genomics is that it may eventually lead to personalized medicine, where patients can take genotype specific drugs for medical treatments.

Personal genomics is not a single individual's vision or invention. Many researchers for decades anticipated this biological branch will eventually arrive with minimum cost of genotyping. Due to the advent of cheap and fast sequencers, full genome personal genomics is becoming a reality. However, there have been active early proponents of personal genomics projects such as George Church in Harvard Medical School.

Genomics used to mean academic research on consensus genomes which have been assembled from many different individuals of a particular species. The personal genomics changes this into customized bioinformatic discovery on individuals.

Use of personal genomics in predictive medicine

Predictive medicine is the use of the information produced by personal genomics techniques when deciding what medical treatments are appropriate for a particular individual.

An example of the use of predictive medicine is pharmacogenomics, in which genetic information can be used to select the most appropriate drug to prescribe to a patient. The drug should be chosen to maximize the probability of obtaining the desired result in the

patient and minimize the probability that the patient will experience side effects. It is hoped that genetic information will allow physicians to tailor therapy to a given patient, in order to increase drug efficacy and minimize side effects. There are only a few examples in which this information is currently useful in clinical practice, but it is anticipated that tailored therapy will emerge rapidly as researchers validate the clinical utility of different pharmacogenomic markers.

Another area in which there is great interest is disease risk prediction based on genetic markers. Researchers in this area have generated a great deal of information through the use of genome-wide association studies. While there is hope that risk information will be useful in providing predictive medicine, most common medical conditions are multifactorial and the actual risk to the individual depends on both genetic and environmental components, both of which are not completely understood at present. Therefore, the clinical utility of personal genomic information is currently limited. It is hoped that with further research, an accurate risk profile might enable individuals to take steps to prevent diseases for which they are at increased risk based on genetics.

Cost of sequencing an individual's genome

There is currently great interest in personal genomics. This is being fuelled by the rapid drop in the cost of sequencing a human genome. This drop in cost is due to the continual development of new, faster, cheaper DNA sequencing technologies such as "next generation DNA sequencing" that may provide access to full genome sequencing so that the entire genetic code of an individual can be deduced all at once.

The National Human Genome Research Institute, part of the U.S. National Institute of Health has set a target to be able to sequence a human-sized genome for US$100,000 by 2009 and US$1,000 by 2014. There is a widespread belief that within 10 years the cost of sequencing a human genome will fall to $1,000.

There are 6 billion base pairs in the diploid human genome. Statistical analysis reveals that a coverage of approximately ten times is required to get coverage of both alleles in 90% human genome from 25 base-pair reads with shotgun sequencing. This means a total of 60 billion base pairs that must be sequenced. An Applied Biosystems SOLiD, Illumina or Helicos sequencing machine can sequence 2 to 10 billion base pairs in each $8,000 to $18,000 run. The purchase cost, personnel costs and data processing costs must also be taken into account. Sequencing a human genome therefore costs approximately $300,000 in 2008.

In 2009, Complete Genomics of Mountain View announced that it would provide full genome sequencing for $5,000, from June 2009. This will only be available to institutions, not individuals.

This cost is still too high for governments to introduce programs into health services to sequence the genomes of all individuals in a country. However, it may be viable when it falls below $1,000, and the cost of sequencing a human genome is dropping rapidly. For

example, approximately 1 million babies are born in Canada each year. To sequence all of their genomes would cost approximately $1 billion per year, or just 1% of Canada's total healthcare budget. Given the ethical concerns about presymptomatic genetic testing of minors, it is likely that personal genomics will first be applied to adults who can provide consent to undergo such testing.

In June 2009, Illumina announced that they were launching their own Personal Full Genome Sequencing Service at a depth of 30X for $48,000 per genome.. Only one year later, in 2010, they cut the price 60% to $19,500 . Still too expensive for true commercialization, prices are expected to drop further over the next few years as they realize economies of scale and given the competition with other companies such as Complete Genomics.

Knome's whole genome sequencing approach aims, instead, to read every site in the whole euchromatic portion of a person's genome (roughly 3 billion sites). While significantly more expensive than SNP chip-based genotyping, this approach yields significantly more data, identifying both novel (never-before-seen) and known sequence variants, some of which may be particularly relevant in efforts to understand personal health, as well as ancestry.

Timeline of Personal genomes sequenced

Year	Cost	Personal genomes sequenced	Company
2003	$3,000,000,000	1	Various
2009	$48,000	100	Illumina
2010	$19,500	?	Illumina

Comparative genomics

Comparative genomics analysis is concerned with characterising the differences and similarities between whole genomes. It may be applied to both genomes from individuals from different species or individuals from the same species, generally at lower cost than sequencing from scratch. In personal genomics and personalized medicine, we are concerned with comparing the genomes of different humans. It is likely that many of the techniques which are developed in comparative genomic analysis will be useful in personal genomics and personalized medicine. This includes rare and common Single nucleotide polymorphisms (consisting substituting one base pair by another, for example CATGCCGG to CATGACGG), as well as insertion or deletion of one or many base pairs.

Predictive medicine services already available

At least four companies which offer genome-wide personal genomics services already have gone to market and are selling their services direct to consumer. They are likely to be the first of many. However, the validity of individual risk predictions based on SNPs and the clinical utility of this information is currently questionable.

- deCODEme.com charges $2000 to carry out genotyping of approximately 1 million SNPs and provides risk estimates for 47 diseases as well as ancestry analyses.

- Navigenics, began offering SNP-based genomic risk assessments as of April 2008. Navigenics is medically focused and emphasizes a clinician's and genetic counselor's role in interpreting results. The Health Compass comprehensive genetic test for $999 analyzes your genetic predispositions for a variety of health conditions that meet stringent scientific criteria. Navigenics uses Affymetrix Genome-Wide Human SNP Array 6.0 , which genotypes 900,000 SNPs.

- 23andMe sells mail order kits for SNP genotyping. The $199 kit, with $5.00/month subscription, or $499 without a subscription contains everything a consumer needs to take their own saliva sample. The consumer then mails the sample to 23andMe who carry out microarray analysis on it. This provides genotype information for approximately 1,000,000 SNPs. This information is used to estimate the genetic risk of the consumer for 178 diseases and conditions, as well as ancestry analyses.

- Bioresolve describes a similar service to that of 23andMe; however, the Better Business Bureau gave them an "F" reliability rating.

- Knome provides full genome (98% genome) sequencing services for $39,500 for whole genome sequencing and interpretation for consumers. It’s $29,500 for whole genome sequencing and analysis for researchers depending on their requirements.

- HelloGene and HelloGenome personal genome information services describe genotyping and full genome sequencing launched by Theragen in Korea. HelloGenome is the first commercial whole genome sequencing service in Asia while HelloGene is the first in Korea. HelloGene uses Affymetrix SNP chips while HelloGenome uses Solexa machines.

- Illumina, Oxford Nanopore Technologies, Sequenom, Pacific Biosciences, Complete Genomics and 454 Life Sciences are companies focused on commercializing full genome sequencing but are not involved in the predictive medicine (interpretative) side.

Ethical issues

While personalized medicine will certainly be a great asset to healthcare, it opens up several ethical issues which will need to be thought about carefully. No doubt there will be a huge amount of debate concerning the ethics of personalised medicine in the coming years.

Genetic discrimination is discriminating on the grounds of information obtained from an individual's genome. Genetic non-discrimination laws have been enacted in most US states and, at the federal level, by the Genetic Information Nondiscrimination Act (GINA). The GINA legislation prevents discrimination by health insurers and employers but does not apply to life insurance or long-term care insurance.

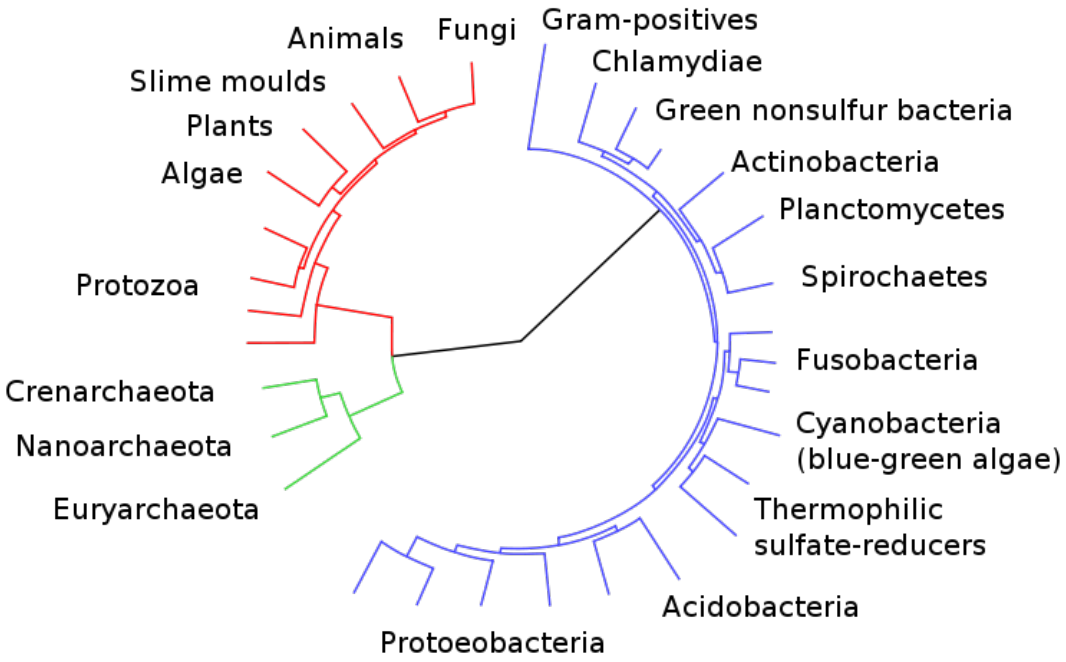
The likelihood of an individual developing breast cancer is affected by which alleles they have of particular genes. Screening can reveal breast cancer in the early stages, allowing it to be successfully treated. 50% of breast cancers occur in the 12% of the population who are at greatest risk. This poses a very difficult question for health services: Is it ethical to deny somebody free screening for a disease if they are genetically at low (but non-zero) risk of developing that disease?

Other issues

Medical genetics will confront the fact that full sequencing of the genome identifies many polymorphisms that are neutral or harmless. This prospect will create uncertainty in the analysis of individual genomes, particularly in the context of clinical care. Czech medical geneticist Eva Machácková writes: "In some cases it is difficult to distinguish if the detected sequence variant is a causal mutation or a neutral (polymorphic) variation without any effect on phenotype. The interpretation of rare sequence variants of unknown significance detected in disease-causing genes becomes an increasingly important problem."

Chapter- 9

Population Genetics



Population genetics is the study of allele frequency distribution and change under the influence of the four main evolutionary processes: natural selection, genetic drift, mutation and gene flow. It also takes into account the factors of population subdivision and population structure. It attempts to explain such phenomena as adaptation and speciation.

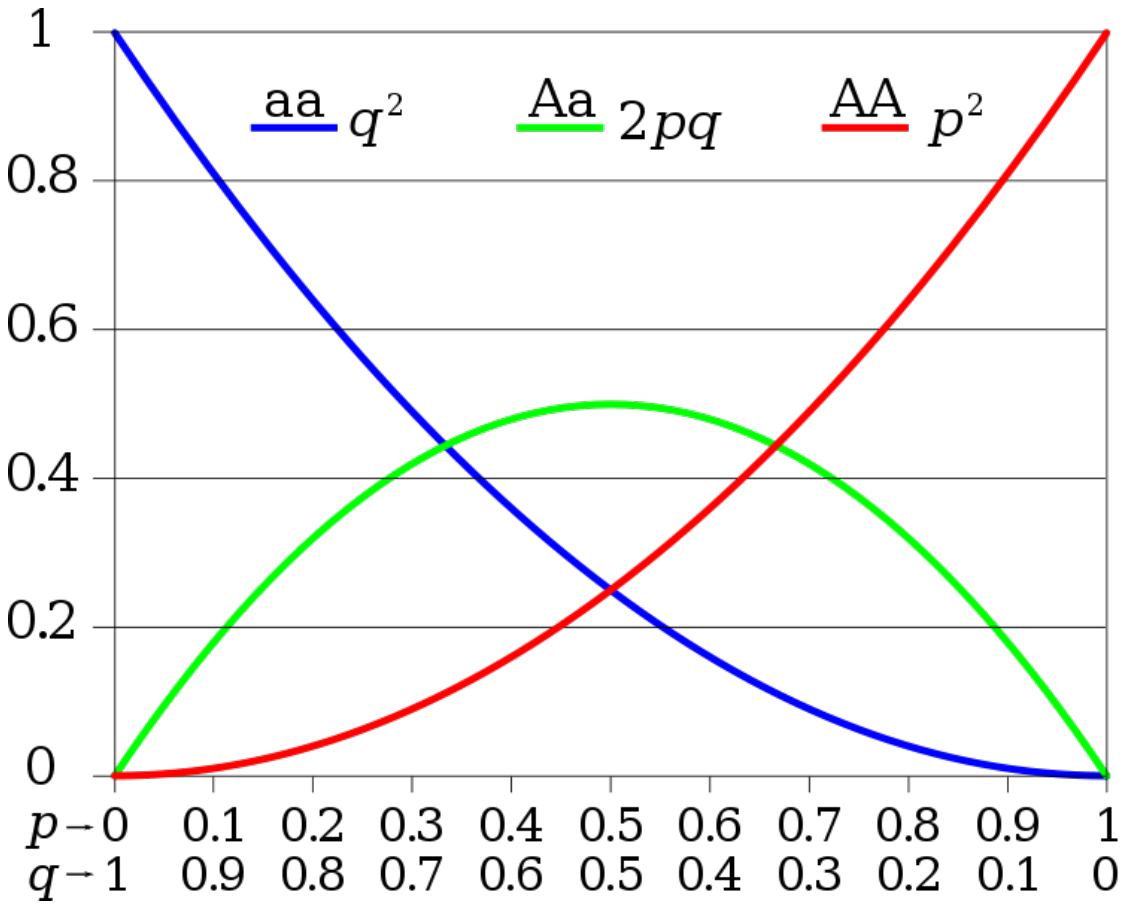
Population genetics was a vital ingredient in the emergence of the modern evolutionary synthesis. Its primary founders were Sewall Wright, J. B. S. Haldane and R. A. Fisher, who also laid the foundations for the related discipline of quantitative genetics.

Fundamentals

Population genetics concerns the genetic constitution of populations and how this constitution changes with time. A population is a set of organisms in which any pair of members can breed together. This implies that all members belong to the same species and live near each other.

For example, all of the moths of the same species living in an isolated forest are a population. A gene in this population may have several alternate forms, which account for variations between the phenotypes of the organisms. An example might be a gene for coloration in moths that has two alleles: black and white. A gene pool is the complete set of alleles for a gene in a single population; the allele frequency for an allele is the fraction of the genes in the pool that is composed of that allele (for example, what fraction of moth coloration genes are the black allele). Evolution occurs when there are changes in the frequencies of alleles within a population of interbreeding organisms; for example, the allele for black color in a population of moths becoming more common.

To understand the mechanisms that cause a population to evolve, it is useful to consider what conditions are required for a population not to evolve. The *Hardy-Weinberg principle* states that the frequencies of alleles (variations in a gene) in a sufficiently large population will remain constant if the only forces acting on that population are the random reshuffling of alleles during the formation of the sperm or egg, and the random combination of the alleles in these sex cells during fertilization. Such a population is said to be in *Hardy-Weinberg equilibrium* as it is not evolving.



Hardy–Weinberg principle for two alleles: the horizontal axis shows the two allele frequencies *p* and *q* and the vertical axis shows the genotype frequencies. Each graph shows one of the three possible genotypes.

Hardy–Weinberg principle

The *Hardy–Weinberg principle* states that both allele and genotype frequencies in a population remain constant—that is, they are in equilibrium—from generation to generation unless specific disturbing influences are introduced. Outside the lab, one or more of these "disturbing influences" are always in effect. Hardy Weinberg equilibrium is impossible in nature. Genetic equilibrium is an ideal state that provides a baseline to measure genetic change against.

Allele frequencies in a population remain static across generations, provided the following conditions are at hand: random mating, no mutation (the alleles don't change), no migration or emigration (no exchange of alleles between populations), infinitely large population size, and no selective pressure for or against any traits.

In the simplest case of a single locus with two alleles: the dominant allele is denoted **A** and the recessive **a** and their frequencies are denoted by *p* and *q*; freq(**A**) = *p*; freq(**a**) = *q*; $p + q = 1$. If the population is in equilibrium, then we will have freq(**AA**) = p^2 for the **AA**

homozygotes in the population, freq(**aa**) = q^2 for the **aa** homozygotes, and freq(**Aa**) = $2pq$ for the heterozygotes.

Based on these equations, useful but difficult-to-measure facts about a population can be determined. For example, a patient's child is a carrier of a recessive mutation that causes cystic fibrosis in homozygous recessive children. The parent wants to know the probability of her grandchildren inheriting the disease. In order to answer this question, the genetic counselor must know the chance that the child will reproduce with a carrier of the recessive mutation. This fact may not be known, but disease frequency is known. We know that the disease is caused by the homozygous recessive genotype; we can use the Hardy–Weinberg principle to work backward from disease occurrence to the frequency of heterozygous recessive individuals.

Scope and theoretical considerations

The mathematics of population genetics were originally developed as part of the modern evolutionary synthesis. According to Beatty (1986), it defines the core of the modern synthesis.

According to Lewontin (1974), the theoretical task for population genetics is a process in two spaces: a "genotypic space" and a "phenotypic space". The challenge of a *complete* theory of population genetics is to provide a set of laws that predictably map a population of genotypes (G_1) to a phenotype space (P_1), where selection takes place, and another set of laws that map the resulting population (P_2) back to genotype space (G_2) where Mendelian genetics can predict the next generation of genotypes, thus completing the cycle. Even leaving aside for the moment the non-Mendelian aspects of molecular genetics, this is clearly a gargantuan task. Visualizing this transformation schematically:

$$G_1 \xrightarrow{T_1} P_1 \xrightarrow{T_2} P_2 \xrightarrow{T_3} G_2 \xrightarrow{T_4} G_1' \ \rightarrow \cdots$$

(adapted from Lewontin 1974, p. 12). XD

T_1 represents the genetic and epigenetic laws, the aspects of functional biology, or development, that transform a genotype into phenotype. We will refer to this as the "genotype-phenotype map". T_2 is the transformation due to natural selection, T_3 are epigenetic relations that predict genotypes based on the selected phenotypes and finally T_4 the rules of Mendelian genetics.

In practice, there are two bodies of evolutionary theory that exist in parallel, traditional population genetics operating in the genotype space and the biometric theory used in plant and animal breeding, operating in phenotype space. The missing part is the mapping between the genotype and phenotype space. This leads to a "sleight of hand" (as Lewontin terms it) whereby variables in the equations of one domain, are considered parameters or *constants*, where, in a full-treatment they would be transformed themselves by the evolutionary process and are in reality *functions* of the state variables in the other domain. The "sleight of hand" is assuming that we know this mapping. Proceeding as if

we do understand it is enough to analyze many cases of interest. For example, if the phenotype is almost one-to-one with genotype (sickle-cell disease) or the time-scale is sufficiently short, the "constants" can be treated as such; however, there are many situations where it is inaccurate.

The four processes

Natural selection

Natural selection is the process by which heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations.

The natural genetic variation within a population of organisms means that some individuals will survive more successfully than others in their current environment. Factors which affect reproductive success are also important, an issue which Charles Darwin developed in his ideas on sexual selection.

Natural selection acts on the phenotype, or the observable characteristics of an organism, but the genetic (heritable) basis of any phenotype which gives a reproductive advantage will become more common in a population. Over time, this process can result in adaptations that specialize organisms for particular ecological niches and may eventually result in the emergence of new species.

Natural selection is one of the cornerstones of modern biology. The term was introduced by Darwin in his groundbreaking 1859 book *On the Origin of Species*, in which natural selection was described by analogy to artificial selection, a process by which animals and plants with traits considered desirable by human breeders are systematically favored for reproduction. The concept of natural selection was originally developed in the absence of a valid theory of heredity; at the time of Darwin's writing, nothing was known of modern genetics. The union of traditional Darwinian evolution with subsequent discoveries in classical and molecular genetics is termed the *modern evolutionary synthesis*. Natural selection remains the primary explanation for adaptive evolution.

Genetic drift

Genetic drift is the change in the relative frequency in which a gene variant (allele) occurs in a population due to random sampling and chance. That is, the alleles in the offspring in the population are a random sample of those in the parents. And chance has a role in determining whether a given individual survives and reproduces. A population's allele frequency is the fraction or percentage of its gene copies compared to the total number of gene alleles that share a particular form.

Genetic drift is an important evolutionary process which leads to changes in allele frequencies over time. It may cause gene variants to disappear completely, and thereby reduce genetic variability. In contrast to natural selection, which makes gene variants

more common or less common depending on their reproductive success, the changes due to genetic drift are not driven by environmental or adaptive pressures, and may be beneficial, neutral, or detrimental to reproductive success.

The effect of genetic drift is larger in small populations, and smaller in large populations. Vigorous debates wage among scientists over the relative importance of genetic drift compared with natural selection. Ronald Fisher held the view that genetic drift plays at the most a minor role in evolution, and this remained the dominant view for several decades. In 1968 Motoo Kimura rekindled the debate with his neutral theory of molecular evolution which claims that most of the changes in the genetic material are caused by genetic drift.

Mutation

Mutations are changes in the DNA sequence of a cell's genome and are caused by radiation, viruses, transposons and mutagenic chemicals, as well as errors that occur during meiosis or DNA replication. Errors are introduced particularly often in the process of DNA replication, in the polymerization of the second strand. These errors can also be induced by the organism itself, by cellular processes such as hypermutation.

Mutations can have an impact on the phenotype of an organism, especially if they occur within the protein coding sequence of a gene. Error rates are usually very low (1 error in every 10 million–100 million bases) due to the "proofreading" ability of DNA polymerases. Without proofreading, error rates are a thousandfold higher. Chemical damage to DNA occurs naturally as well, and cells use DNA repair mechanisms to repair mismatches and breaks in DNA. Nevertheless, the repair sometimes fails to return the DNA to its original sequence.

In organisms that use chromosomal crossover to exchange DNA and recombine genes, errors in alignment during meiosis can also cause mutations. Errors in crossover are especially likely when similar sequences cause partner chromosomes to adopt a mistaken alignment; this makes some regions in genomes more prone to mutating in this way. These errors create large structural changes in DNA sequence—duplications, inversions or deletions of entire regions, or the accidental exchanging of whole parts between different chromosomes (called translocation).

Mutation can result in several different types of change in DNA sequences; these can either have no effect, alter the product of a gene, or prevent the gene from functioning. Studies in the fly *Drosophila melanogaster* suggest that if a mutation changes a protein produced by a gene, this will probably be harmful, with about 70 percent of these mutations having damaging effects, and the remainder being either neutral or weakly beneficial. Due to the damaging effects that mutations can have on cells, organisms have evolved mechanisms such as DNA repair to remove mutations. Therefore, the optimal mutation rate for a species is a trade-off between costs of a high mutation rate, such as deleterious mutations, and the metabolic costs of maintaining systems to reduce the mutation rate, such as DNA repair enzymes. Viruses that use RNA as their genetic

material have rapid mutation rates, which can be an advantage since these viruses will evolve constantly and rapidly, and thus evade the defensive responses of e.g. the human immune system.

Mutations can involve large sections of DNA becoming duplicated, usually through genetic recombination. These duplications are a major source of raw material for evolving new genes, with tens to hundreds of genes duplicated in animal genomes every million years. Most genes belong to larger families of genes of shared ancestry. Novel genes are produced by several methods, commonly through the duplication and mutation of an ancestral gene, or by recombining parts of different genes to form new combinations with new functions.

Here, domains act as modules, each with a particular and independent function, that can be mixed together to produce genes encoding new proteins with novel properties. For example, the human eye uses four genes to make structures that sense light: three for color vision and one for night vision; all four arose from a single ancestral gene. Another advantage of duplicating a gene (or even an entire genome) is that this increases redundancy; this allows one gene in the pair to acquire a new function while the other copy performs the original function. Other types of mutation occasionally create new genes from previously noncoding DNA.

Gene flow

Gene flow is the exchange of genes between populations, which are usually of the same species. Examples of gene flow within a species include the migration and then breeding of organisms, or the exchange of pollen. Gene transfer between species includes the formation of hybrid organisms and horizontal gene transfer.

Migration into or out of a population can change allele frequencies, as well as introducing genetic variation into a population. Immigration may add new genetic material to the established gene pool of a population. Conversely, emigration may remove genetic material. As barriers to reproduction between two diverging populations are required for the populations to become new species, gene flow may slow this process by spreading genetic differences between the populations. Gene flow is hindered by mountain ranges, oceans and deserts or even man-made structures such as the Great Wall of China, which has hindered the flow of plant genes.

Depending on how far two species have diverged since their most recent common ancestor, it may still be possible for them to produce offspring, as with horses and donkeys mating to produce mules. Such hybrids are generally infertile, due to the two different sets of chromosomes being unable to pair up during meiosis. In this case, closely related species may regularly interbreed, but hybrids will be selected against and the species will remain distinct. However, viable hybrids are occasionally formed and these new species can either have properties intermediate between their parent species, or possess a totally new phenotype. The importance of hybridization in creating new species

of animals is unclear, although cases have been seen in many types of animals, with the gray tree frog being a particularly well-studied example.

Hybridization is, however, an important means of speciation in plants, since polyploidy (having more than two copies of each chromosome) is tolerated in plants more readily than in animals. Polyploidy is important in hybrids as it allows reproduction, with the two different sets of chromosomes each being able to pair with an identical partner during meiosis. Polyploids also have more genetic diversity, which allows them to avoid inbreeding depression in small populations.

Horizontal gene transfer is the transfer of genetic material from one organism to another organism that is not its offspring; this is most common among bacteria. In medicine, this contributes to the spread of antibiotic resistance, as when one bacteria acquires resistance genes it can rapidly transfer them to other species. Horizontal transfer of genes from bacteria to eukaryotes such as the yeast *Saccharomyces cerevisiae* and the adzuki bean beetle *Callosobruchus chinensis* may also have occurred. An example of larger-scale transfers are the eukaryotic bdelloid rotifers, which appear to have received a range of genes from bacteria, fungi, and plants. Viruses can also carry DNA between organisms, allowing transfer of genes even across biological domains. Large-scale gene transfer has also occurred between the ancestors of eukaryotic cells and prokaryotes, during the acquisition of chloroplasts and mitochondria.

Gene flow is the transfer of alleles from one population to another.

Migration into or out of a population may be responsible for a marked change in allele frequencies. Immigration may also result in the addition of new genetic variants to the established gene pool of a particular species or population.

There are a number of factors that affect the rate of gene flow between different populations. One of the most significant factors is mobility, as greater mobility of an individual tends to give it greater migratory potential. Animals tend to be more mobile than plants, although pollen and seeds may be carried great distances by animals or wind.

Maintained gene flow between two populations can also lead to a combination of the two gene pools, reducing the genetic variation between the two groups. It is for this reason that gene flow strongly acts against speciation, by recombining the gene pools of the groups, and thus, repairing the developing differences in genetic variation that would have led to full speciation and creation of daughter species.

For example, if a species of grass grows on both sides of a highway, pollen is likely to be transported from one side to the other and vice versa. If this pollen is able to fertilise the plant where it ends up and produce viable offspring, then the alleles in the pollen have effectively been able to move from the population on one side of the highway to the other.

Genetic structure

Because of physical barriers to migration, along with limited vagility, and natal philopatry, natural populations are rarely panmictic (Buston *et al.*, 2007). There is usually a geographic range within which individuals are more closely related to one another than those randomly selected from the general population. This is described as the extent to which a population is genetically structured (Repaci *et al.*, 2007).

Microbial population genetics

Microbial population genetics is a rapidly advancing field of investigation with relevance to many other theoretical and applied areas of scientific investigations. The population genetics of microorganisms lays the foundations for tracking the origin and evolution of antibiotic resistance and deadly infectious pathogens. Population genetics of microorganisms is also an essential factor for devising strategies for the conservation and better utilization of beneficial microbes (Xu, 2010).

History

Biston betularia f. typica is the white-bodied form of the peppered moth.

Biston betularia f. carbonaria is the black-bodied form of the peppered moth.

Population genetics

Population genetics was developed as a reconciliation of the Mendelian and biometrician models. A key step was the work of the British biologist and statistician R.A. Fisher. In a series of papers starting in 1918 and culminating in his 1930 book *The Genetical Theory of Natural Selection*, Fisher showed that the continuous variation measured by the biometricians could be produced by the combined action of many discrete genes, and that natural selection could change gene frequencies in a population, resulting in evolution (though lacking the knowledge of what an actual gene was at this time, it should be said in this sense he understood phenotypic trait frequency, rather than specifically identifiable gene frequency). In a series of papers beginning in 1924, another British geneticist, J.B.S. Haldane, applied statistical analysis to real-world examples of natural selection, such as the evolution of industrial melanism in peppered moths, and showed that natural selection worked at an even faster rate than Fisher assumed.

The American biologist Sewall Wright, who had a background in animal breeding experiments, focused on combinations of interacting genes, and the effects of inbreeding on small, relatively isolated populations that exhibited genetic drift. In 1932, Wright introduced the concept of an adaptive landscape and argued that genetic drift and inbreeding could drive a small, isolated sub-population away from an adaptive peak, allowing natural selection to drive it towards different adaptive peaks. Fisher and Wright had some fundamental disagreements and a controversy about the relative roles of selection and drift continued for much of the century between the Americans and the

British. The Frenchman Gustave Malécot was also important early in the development of the discipline.

The work of Fisher, Haldane and Wright founded the discipline of *population genetics*. This integrated natural selection with Mendelian genetics, which was the critical first step in developing a unified theory of how evolution worked.

John Maynard Smith was Haldane's pupil, whilst W.D. Hamilton was heavily influenced by the writings of Fisher. The American George R. Price worked with both Hamilton and Maynard Smith. American Richard Lewontin and Japanese Motoo Kimura were heavily influenced by Wright.

Modern evolutionary synthesis

In the first few decades of the 20th century, most field naturalists continued to believe that Lamarckian and orthogenic mechanisms of evolution provided the best explanation for the complexity they observed in the living world. However, as the field of genetics continued to develop, those views became less tenable. Theodosius Dobzhansky, a postdoctoral worker in T. H. Morgan's lab, had been influenced by the work on genetic diversity by Russian geneticists such as Sergei Chetverikov. He helped to bridge the divide between the foundations of microevolution developed by the population geneticists and the patterns of macroevolution observed by field biologists, with his 1937 book *Genetics and the Origin of Species*.

Dobzhansky examined the genetic diversity of wild populations and showed that, contrary to the assumptions of the population geneticists, these populations had large amounts of genetic diversity, with marked differences between sub-populations. The book also took the highly mathematical work of the population geneticists and put it into a more accessible form. In Great Britain E.B. Ford, the pioneer of ecological genetics, continued throughout the 1930s and 1940s to demonstrate the power of selection due to ecological factors including the ability to maintain genetic diversity through genetic polymorphisms such as human blood types. Ford's work would contribute to a shift in emphasis during the course of the modern synthesis towards natural selection over genetic drift.

Chapter- 10

Allele and Allele Frequency

Allele

An **allele** is one of two or more forms of the DNA sequence of a particular gene. The word is a short form of allelomorph ('other form'), which was used in the early days of genetics to describe variant forms of a gene detected as different phenotypes.

Each gene can have different alleles. Sometimes, different DNA sequences (alleles) can result in different traits, such as color. Sometimes, different DNA sequences (alleles) will have the same result in the expression of a gene.

Most multicellular organisms have two sets of chromosomes, that is, they are diploid. These chromosomes are referred to as homologous chromosomes. Diploid organisms have one copy of each gene (and one allele) on each chromosome. If both alleles are the same, they are homozygotes. If the alleles are different, they are heterozygotes.

A population or species of organisms typically includes multiple alleles at each locus among various individuals. Allelic variation at a locus is measurable as the number of alleles (polymorphism) present, or the proportion of heterozygotes (heterozygosity) in the population.

For example, at the gene locus for ABO blood type proteins in humans, classical genetics recognizes three alleles, I^A, I^B, and I^O, that determines compatibility of blood transfusions. Any individual has one of six possible genotypes (AA, AO, BB, BO, AB, and OO) that produce one of four possible phenotypes: "A" (produced by AA homozygous and AO heterozygous genotypes), "B" (produced by BB homozygous and BO heterozygous genotypes), "AB" heterozygotes, and "O" homozygotes. It is now appreciated that each of the A, B, and O alleles is actually a class of multiple alleles with different DNA sequences that produce proteins with identical properties: more than 70 alleles are known at the ABO locus. An individual with "Type A" blood may be a AO heterozygote, an AA homozygote, or an A'A heterozygote with two different 'A' alleles.

Dominant and recessive alleles

In many cases, genotypic interactions between the two alleles at a locus can be described as dominant or recessive, according to which of the two homozygous genotype the phenotype of the heterozygote most resembles. Where the heterozygote is indistinguishable from one of the homozygotes, the allele involved is said to be dominant to the other, which is said to be recessive to the former. The degree and pattern of dominance varies among loci.

The term "wild type" allele is sometimes used to describe an allele that is thought to contribute to the typical phenotypic character as seen in "wild" populations of organisms, such as fruit flies (*Drosophila melanogaster*). Such a "wild type" allele was historically regarded as dominant, common, and "normal", in contrast to "mutant" alleles regarded as recessive, rare, and frequently deleterious. It was commonly thought that most individuals were homozygous for the "wild type" allele at most gene loci, and that any alternative 'mutant' allele was found in homozygous form in a small minority of "affected" individuals, often as genetic diseases, and more frequently in heterozygous form in "carriers" for the mutant allele. It is now appreciated that most or all gene loci are highly polymorphic, with multiple alleles, whose frequencies vary from population to population, and that a great deal of genetic variation is hidden in the form of alleles that do not produce obvious phenotypic differences.

Allele and genotype frequencies

The frequency of alleles in a population can be used to predict the frequencies of the corresponding genotypes. For a simple model, with two alleles:

$$p + q = 1$$
$$p^2 + 2pq + q^2 = 1$$

where p is the frequency of one allele and q is the frequency of the alternative allele, which necessarily sum to unity. Then, p^2 is the fraction of the population homozygous for the first allele, $2pq$ is the fraction of heterozygotes, and q^2 is the fraction homozygous for the alternative allele. If the first allele is dominant to the second, then the fraction of the population that will show the dominant phenotype is $p^2 + 2pq$, and the fraction with the recessive phenotype is q^2.

With three alleles:

$$p + q + r = 1_{\text{and}}$$
$$p^2 + 2pq + 2pr + q^2 + 2qr + r^2 = 1$$

In the case of multiple alleles at a diploid locus, the number of possible genotypes (G) with a number of alleles (a) is given by the expression:

$$G = \frac{a(a+1)}{2}$$

Allelic variation in genetic disorders

A number of genetic disorders are caused when an individual inherits two recessive alleles for a single-gene trait. Recessive genetic disorders include Albinism, Cystic Fibrosis, Galactosemia, Phenylketonuria (PKU), and Tay-Sachs Disease. Other disorders are also due to recessive alleles, but because the gene locus is located on the X chromosome, so that males have only one copy (that is, they are hemizygous), they are more frequent in males than in females. Examples include red-green color blindness and Fragile X syndrome.

Allele frequency

Allele frequency is the proportion of all copies of a gene that is made up of a particular gene variant (allele). In other words, it is the number of copies of a particular allele divided by the number of copies of all alleles at the genetic place (locus) in a population. It can be expressed for example as a percentage. In population genetics, allele frequencies are used to depict the amount of genetic diversity at the individual, population, and species level. It is also the relative proportion of all alleles of a gene that are of a designed type.

Given the following:

1. a particular locus on a chromosome and the gene occupying that locus
2. a population of *N* individuals carrying *n* loci in each of their somatic cells (e.g. two loci in the cells of diploid species, which contain two sets of chromosomes)
3. different alleles of the gene exist
4. one allele exists in *a* copies

then the allele frequency is the fraction or percentage of all the occurrences of that locus that is occupied by a given allele and the frequency of one of the alleles is *a/(n*N)*.

For example, if the frequency of an allele is 20% in a given population, then among population members, one in five chromosomes will carry that allele. Four out of five will be occupied by other variant(s) of the gene. Note that for diploid genes the fraction of individuals that carry this allele may be nearly *two in five*. If the allele distributes randomly, then the binomial theorem will apply: 32% of the population will be heterozygous for the allele (i.e. carry one copy of that allele and one copy of another in each somatic cell) and 4% will be homozygous (carrying two copies of the allele). Together, this means that 36% of diploid individuals would be expected to carry an allele

that has a frequency of 20%. However, alleles distribute randomly only under certain assumptions, including the absence of selection. When these conditions apply, a population is said to be in Hardy–Weinberg equilibrium.

The frequencies of all the alleles of a given gene often are graphed together as an *allele frequency distribution histogram*, or *allele frequency spectrum*. Population genetics studies the different "forces" that might lead to changes in the distribution and frequencies of alleles—in other words, to evolution. Besides selection, these forces include genetic drift, mutation and migration.

Calculation of allele frequencies from genotype frequencies

If *f*(*AA*), *f*(*Aa*), and *f*(*aa*) are the frequencies of the three genotypes at a locus with two alleles, then the frequency *p* of the A-allele and the frequency *q* of the a-allele are obtained by counting alleles. Because each homozygote AA consists only of A-alleles, and because half of the alleles of each heterozygote Aa are A-alleles, the total frequency *p* of A-alleles in the population is calculated as

$$p = f(\mathbf{AA}) + \frac{1}{2} f(\mathbf{Aa}) = \text{frequency of A}$$

Similarly, the frequency *q* of the a allele is given by

$$q = f(\mathbf{aa}) + \frac{1}{2} f(\mathbf{Aa}) = \text{frequency of a}$$

It would be expected that *p* and *q* sum to 1, since they are the frequencies of the only two alleles present. Indeed they do:

$$p + q = f(\mathbf{AA}) + f(\mathbf{aa}) + f(\mathbf{Aa}) = 1$$

and from this we get:

$q = 1 - p$ and $p = 1 - q$

If there are more than two different allelic forms, the frequency for each allele is simply the frequency of its homozygote plus half the sum of the frequencies for all the heterozygotes in which it appears. Allele frequency can always be calculated from genotype frequency, whereas the reverse requires that the Hardy–Weinberg conditions of random mating apply. This is partly due to the *three* genotype frequencies and the *two* allele frequencies. It is easier to reduce from three to two.

An example population

Consider a population of ten individuals and a given locus with two possible alleles, *A* and *a*. Suppose that the genotypes of the individuals are as follows:

AA, *Aa*, *AA*, *aa*, *Aa*, *AA*, *AA*, *Aa*, *Aa*, and *AA*

Then the allele frequencies of allele *A* and allele *a* are:

$$p = prob_A = \frac{2+1+2+0+1+2+2+1+1+2}{20} = 0.7$$
$$q = prob_a = \frac{0+1+0+2+1+0+0+1+1+0}{20} = 0.3$$

so if an individual is chosen at random there is a 70% chance it will carry the *A* allele, and a 30% chance it will have the *a* allele.

The effect of mutation

Let ú be the mutation rate from allele *A* to some other allele *a* (the probability that a copy of gene *A* will become *a* during the DNA replication preceding meiosis). If p_t is the frequency of the *A* allele in generation *t*, then $q_t = 1 - p_t$ is the frequency of the *a* allele in generation *t*, and if there are no other causes of gene frequency change (no natural selection, for example), then the change in allele frequency in one generation is

$$\Delta p = p_t - p_{t-1} = (p_{t-1} - úp_{t-1}) - p_{t-1} = -úp_{t-1}$$

where p_{t-1} is the frequency of the preceding generation. This tells us that the frequency of *A* decreases (and the frequency of *a* increases) by an amount that is proportional to the mutation rate ú and to the proportion *p* of all the genes that are still available to mutate. Thus Δ*p* gets smaller as the frequency of *p* itself decreases, because there are fewer and fewer *A* alleles to mutate into *a* alleles. We can make an approximation that, after *n* generations of mutation,

$$p_n = p_0 e^{-nú}$$

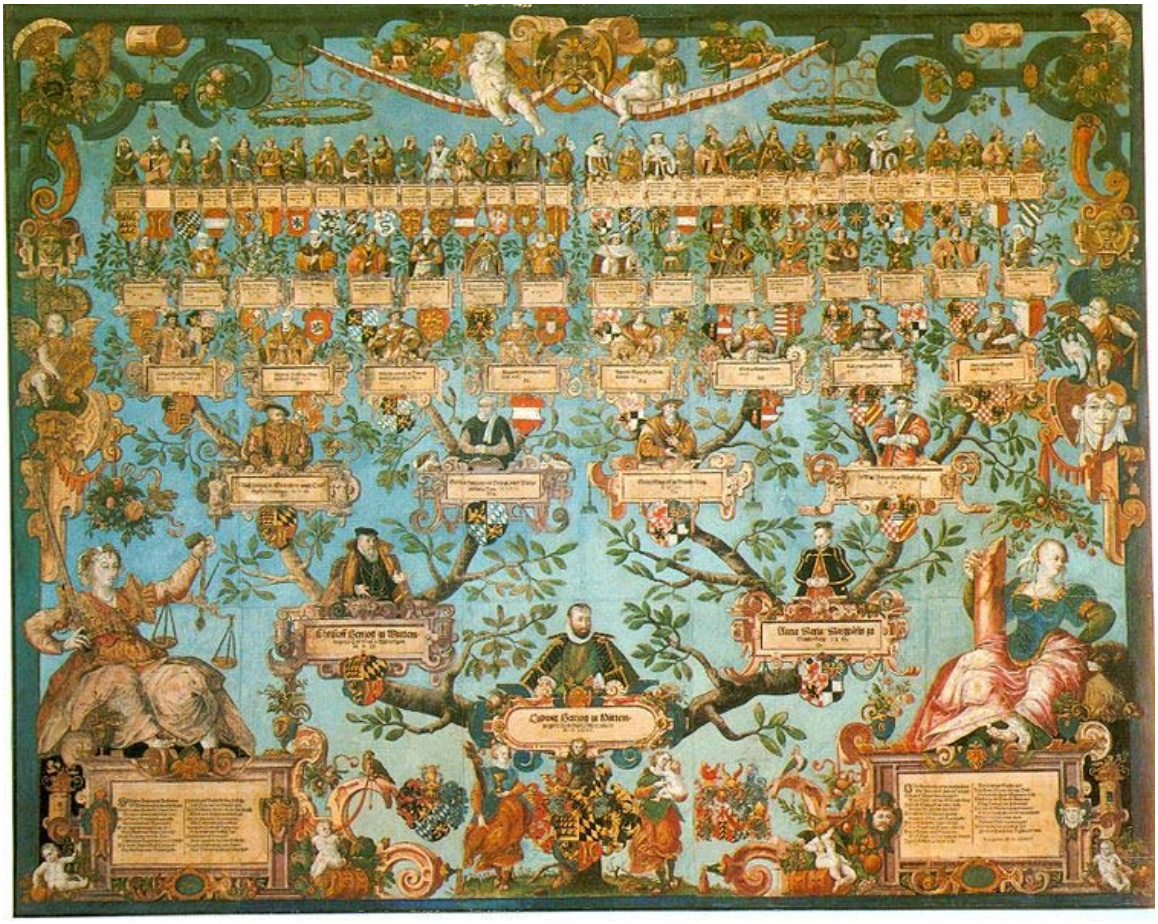
Chapter- 11

Genealogy

Genealogy (from Greek: γενεά, *genea*, "generation"; and λόγος, *logos*, "knowledge") is the study of families and the tracing of their lineages and history. Genealogists use oral traditions, historical records, genetic analysis, and other records to obtain information about a family and to demonstrate kinship and pedigrees of its members. The results are often displayed in charts or written as narratives.

The pursuit of family history tends to be shaped by several motivations, including the desire to carve out a place for one's family in the larger historical picture, a sense of responsibility to preserve the past for future generations, and a sense of self-satisfaction in accurate storytelling.

Some scholars differentiate between genealogy and family history, limiting genealogy to an account of kinship, while using "family history" to denote the provision of additional details about lives and historical context.



The family tree of Herzog Ludwig I of Württemberg (ruled 1568–1593)

Overview

Hobbyist genealogists typically pursue their own ancestry and that of their spouses. Professional genealogists may also conduct research for others, publish books on genealogical methods, teach, or work for companies that provide software or online databases. Both try to understand not just where and when people lived, but also their lifestyles, biographies, and motivations. This often requires—or leads to—knowledge of antiquated laws, old political boundaries, migration trends, and historical social conditions.

Genealogists sometimes specialize in a particular group, e.g. a Scottish clan; a particular surname, such as in a one-name study; a small community, e.g. a single village or parish, such as in a one-place study; or a particular, often famous, person. Bloodlines of Salem is an example of a specialized family-history group. It welcomes members who are able to prove descent from a participant of the Salem Witch Trials or who choose simply to support the group.

Genealogists and family historians often join family history societies, where novices can learn from more experienced researchers. Such societies may also index records to make them more accessible, and engage in advocacy and other efforts to preserve public records and cemeteries.

Historical background

Historically, in Western societies the focus of genealogy was on the kinship and descent of rulers and nobles, often arguing or demonstrating the legitimacy of claims to wealth and power. The term often overlapped with heraldry, in which the ancestry of royalty was reflected in their coats of arms. Many claimed noble ancestries are considered fabrications by modern scholars, such as the Anglo-Saxon chronicles that traced the ancestry of several English kings to the god Woden.

Genealogical research in the United States was first systematized in the early 19th century, especially by John Farmer (1789–1838). Before Farmer's efforts, tracing one's genealogy was seen as an attempt by colonists to secure a measure of social standing within the British Empire, an aim that was counter to the new republic's egalitarian, future-oriented ethos. As Fourth of July celebrations commemorating the Founding Fathers and the heroes of the Revolutionary War became increasingly popular, however, the pursuit of 'antiquarianism,' which focused on local history, became acceptable as a way to honor the achievements of early Americans. Farmer capitalized on the acceptability of antiquarianism to frame genealogy within the early republic's ideological framework of pride in one's American ancestors. He corresponded with other antiquarians in New England, where antiquarianism and genealogy were well established, and became a coordinator, booster, and contributor to the growing movement. In the 1820s, he and fellow antiquarians began to produce genealogical and antiquarian tracts in earnest, slowly gaining a devoted audience among the American people. Though Farmer died in 1839, his efforts led to the creation of the New England Historic Genealogical Society, which publishes the *New England Historical and Genealogical Register.* The society is one of New England's oldest and most prominent organizations dedicated to the acquisition, preservation, and dissemination of public records and private monuments that would otherwise have decayed and been forgotten.

The Genealogical Society of Utah, founded in 1894, later became the Family History Department of the Mormon church. The department's research facility, the Family History Library, which has developed the most extensive genealogical record-gathering program in the world, was established to assist in tracing family lineages for special religious ceremonies that Mormons believe will seal family units together for eternity. Mormons believe that this fulfilled a biblical prophecy stating that the prophet Elijah would return to 'turn the heart of the fathers to the children, and the heart of the children to their fathers.'

In modern times, genealogy became more widespread, with commoners as well as nobility researching and maintaining their family trees. Genealogy received a boost in the late 1970s with the premiere of the television adaptation of Alex Haley's account of his

family line, *Roots: The Saga of an American Family*, With the advent of the Internet, the number of resources readily accessible by genealogists has vastly increased, resulting in an explosion of interest in the topic. According to some sources, genealogy is one of the most popular topics on the Internet. The Internet has become not only a major source of data for genealogists, but also of education and communication.

Genealogical research process

Genealogical research is a complex process that uses historical records and sometimes genetic analysis to demonstrate kinship. Reliable conclusions are based on the quality of sources, ideally original records, the information within those sources, ideally primary or firsthand information, and the evidence that can be drawn, directly or indirectly, from that information. In many instances, genealogists must skillfully assemble indirect or circumstantial evidence to build a case for identity and kinship. All evidence and conclusions, together with the documentation that supports them, is then assembled to create a cohesive "genealogy" or "family history". Historical, social, and family context is essential to achieving correct identification of individuals and relationships. Source citation is also important when conducting genealogical research.

Genealogists begin their research by collecting family documents and stories. This creates a foundation for documentary research, which involves examining and evaluating historical records for evidence about ancestors and other relatives, their kinship ties, and the events that occurred in their lives. As a rule, genealogists begin with the present and work backward in time.

To keep track of collected material, family group sheets and pedigree charts are used. Formerly handwritten, these can now be generated by genealogical software.

Genetic analysis

Because a person's DNA contains information that has been passed down relatively unchanged from early ancestors, analysis of DNA is sometimes used for genealogical research. Two DNA types are of particular interest: mitochondrial DNA that we all possess and that is passed down with only minor mutations through the matrilineal (direct female) line; and the Y-chromosome, present only in males, which is passed down with only minor mutations through the patrilineal (direct male) line.

A genealogical DNA test allows two individuals to find the probability that they are, or are not, related within an estimated number of generations. Individual genetic test results are collected in databases to match people descended from a relatively recent common ancestor. See, for example, the Molecular Genealogy Research Project. These tests are limited to either the patrilineal or the matrilineal line.

Data sharing among researchers

Most genealogy software programs can export information about persons and their relationships in a standardized format called "GEDCOM" In that format it can be shared with other genealogists, added to online databases, or converted into family web sites. Social networking service (SNS) websites allow genealogists to share data and build their family trees online. Members can upload their family trees and contact other family historians to fill in gaps in their research.

Volunteerism

Volunteer efforts figure prominently in genealogy. These range from the extremely informal to the highly organized.

On the informal side are the many popular and useful message boards such as Rootschat and mailing lists on particular surnames, regions, and other topics. These forums can be used to try to find relatives, request record lookups, obtain research advice, and much more.

Many genealogists participate in loosely organized projects, both online and off. These collaborations take numerous forms. Some projects prepare name indexes for records, such as probate cases, and publish the indexes, either off- or online. These indexes can be used as finding aids to locate original records. Other projects transcribe or abstract records. Offering record lookups for particular geographic areas is another common service. Volunteers, such as those involved in Random Acts of Genealogical Kindness (RAOGK), do record lookups in their home areas for researchers who are unable to travel.

Those looking for a structured volunteer environment can join one of thousands of genealogical societies worldwide. Most societies have a unique area of focus, such as a particular surname, ethnicity, geographic area, or descendancy from participants in a given historical event. Genealogical societies are almost exclusively staffed by volunteers and may offer a broad range of services, including maintaining libraries for members' use, publishing newsletters, providing research assistance to the public, offering classes or seminars, and organizing record preservation or transcription projects.

Records in genealogical research

Genealogists use a wide variety of records in their research. To effectively conduct genealogical research, it is important to understand how the records were created, what information is included in them, and how and where to access them.

Records that are used in genealogy research include:

- Vital records
    - Birth records

    - Death records
    - Marriage and divorce records
- Adoption records
- Biographies and biographical profiles (e.g. *Who's Who*)
- Census records
- Church records
    - Baptism or christening
    - Confirmation
    - Bar or bat mitzvah
    - Marriage
    - Funeral or death
    - Membership
- City directories and telephone directories
- Coroner's reports
- Court records
    - Criminal records
    - Civil records
- Diaries, personal letters and family Bibles
- Emigration, immigration and naturalization records
- Hereditary & lineage organization records, e.g. Daughters of the American Revolution records
- Land and property records, deeds
- Medical records
- Military and conscription records
- Newspaper articles
- Obituaries
- Occupational records
- Oral histories
- Passports
- Photographs
- Poorhouse, workhouse, almshouse, and asylum records
- School and alumni association records
- Ship passenger lists
- Social Security (within the US) and pension records
- Tax records
- Tombstones, cemetery records, and funeral home records
- Voter registration records
- Wills and probate records

To keep track of their citizens, governments began keeping records of persons who were neither royalty nor nobility. In England and Germany, for example, such record keeping started with parish registers in the 16th century. As more of the population was recorded, there were sufficient records to follow a family. Major life events, such as births, marriages, and deaths, were often documented with a license, permit, or report. Genealogists locate these records in local, regional or national offices or archives and extract information about family relationships and recreate timelines of persons' lives.

In China, India and other Asian countries, genealogy books are used to record the names, occupations, and other information about family members, with some books dating back hundreds or even thousands of years. In the eastern Indian state of Bihar, there is a written tradition of genealogical records among Maithil Brahmins and Karna Kayasthas called "Panjis", dating to the 12th century CE. Even today these records are consulted prior to marriages.

In Ireland, genealogical records were recorded by professional families of *senchaidh* (historians) until as late as the mid-17th century, when Gaelic civilization died out. Perhaps the most outstanding example of this genre is Leabhar na nGenealach/The Great Book of Irish Genealogies, by Dubhaltach MacFhirbhisigh (d. 1671), published in 2004.

LDS collections

The Family History Library, operated by the LDS Church, is the world's largest library dedicated to genealogical research

The Church of Jesus Christ of Latter-day Saints (LDS) has engaged in large-scale microfilming of records of genealogical value. Their Family History Library in Salt Lake City, Utah, houses over 2 million microfiche and microfilms of genealogically relevant material, which are also available for on-site research at over 4500 Family History Centers worldwide.

The LDS church has also compiled indexes of the submissions of its members, resulting in several large databases: the International Genealogical Index, or IGI, which includes both data extracted from filmed civil and ecclesiastic records from various worldwide locales and member-submitted information; the Ancestral File, or AF, which includes the contributions of church members; and the Pedigree Resource File, or PRF, compiled from member and non-member submissions. The IGI contains indexes to millions of records of individuals who lived between 1500 and 1900, primarily in the United States, Canada and Europe. Although independent of the IGI, the AF and PRF often contain duplications of IGI records. All three of these indexes are available free on their website, FamilySearch. FamilySearch also includes an 1880 United States federal census index, an 1881 British census index, an 1881 Canadian census index, and the U.S. Social Security Death Index, as well as research guides and genealogical word lists.

Types of genealogical information

Genealogists who seek to reconstruct the lives of each ancestor consider all historical information to be "genealogical" information. Traditionally, the basic information needed to ensure correct identification of each person are place names, occupations, family names, first names, and dates. However, modern genealogists greatly expand this list, recognizing the need to place this information in its historical context in order to properly evaluate genealogical evidence and distinguish between same-name individuals. A great deal of information is available for British ancestry with growing resources for other ethnic groups.

Family names

Family names are simultaneously one of the most important pieces of genealogical information, and a source of significant confusion for researchers.

In many cultures, the name of a person refers to the family to which he or she belongs. This is called the *family name*, *surname*, or *last name*. Patronymics are names that identify an individual based on the father's name. For example, Marga Olafsdottir is Marga, daughter of Olaf, and Olaf Thorsson is Olaf, son of Thor. Many cultures used patronymics before surnames were adopted or came into use. The Dutch in New York, for example, used the patronymic system of names until 1687 when the advent of English rule mandated surname usage. In Iceland, patronymics are used by a majority of the population. In Denmark and Norway patronymics and farm names were generally in use through the 19th century and beyond, though surnames began to come into fashion toward the end of the 19th century in some parts of the country. Not until 1856 in Denmark and 1923 in Norway were there laws requiring surnames.

The transmission of names across generations, marriages and other relationships, and immigration may cause difficulty in genealogical research. For instance, women in many cultures have routinely used their spouse's surnames. When a woman remarried, she may have changed her name and the names of her children; only her name; or changed no names. Her birth name (maiden name) may be reflected in her children's middle names;

her own middle name; or dropped entirely. Children may sometimes assume stepparent, foster parent, or adoptive parent names. Because official records may reflect many kinds of surname change, without explaining the underlying reason for the change, the correct identification of a person recorded identified with more than one name is challenging. Immigrants often Americanized their names.

Surname data may be found in trade directories, census returns, birth, death, and marriage records.

Given names

Genealogical data regarding given names (first names) is subject to many of the same problems as are family names and place names. Additionally, the use of nicknames is very common. For example Beth, Lizzie or Betty are all common for Elizabeth, and Jack, John and Jonathan may be interchanged.

Middle names provide additional information. Middle names may be inherited, follow naming customs, or be treated as part of the family name. For instance, in some Latin cultures, both the mother's family name and the father's family name are used by the children.

Historically, naming traditions existed in some places and cultures. Even in areas that tended to use naming conventions, however, they were by no means universal. Families may have used them some of the time, among some of their children, or not at all. A pattern might also be broken to name a newborn after a recently deceased sibling, aunt or uncle.

An example of a naming tradition from England, Scotland and Ireland:

Child	**Namesake**
1st son	paternal grandfather
2nd son	maternal grandfather
3rd son	father
4th son	father's oldest brother
1st daughter	maternal grandmother
2nd daughter	paternal grandmother
3rd daughter	mother
4th daughter	mother's oldest sister

Another example is in some areas of Germany, where siblings were given the same first name, often of a favourite saint or local nobility, but different second names by which they were known (*Rufname*). If a child died, the next child of the same gender that was born may have been given the same name. It is not uncommon that a list of a particular couple's children will show one or two names repeated.

Personal names have periods of popularity, so it is not uncommon to find many similarly named people in a generation, and even similarly named families; e.g., "William and Mary and their children David, Mary, and John".

Many names may be identified strongly with a particular gender; e.g., William for boys, and Mary for girls. Others may be ambiguous, e.g., Lee, or have only slightly variant spellings based on gender, e.g., Frances (usually female) and Francis (usually male).

Place names

While the locations of ancestors' residences and life events are core elements of the genealogist's quest, they can often be confusing. Place names may be subject to variant spellings by partially literate scribes. Locations may have identical or very similar names. For example, the village name Brockton occurs six times in the border area between the English counties of Shropshire and Staffordshire. Shifts in political borders must also be understood. Parish, county and national borders have frequently been modified. Old records may contain references to farms and villages that have ceased to exist. When working with older records from Poland, where borders and place names have changed frequently in past centuries, a source with maps and sample records such as *A Translation Guide to 19th-Century Polish-Language Civil-Registration Documents* can be invaluable.

Available sources may include vital records (civil or church registration), censuses, and tax assessments. Oral tradition is also an important source, although it must be used with caution. When no source information is available for a location, circumstantial evidence may provide a probable answer based on a person's or a family's place of residence at the time of the event.

Maps and gazetteers are important sources for understanding the places researched. They show the relationship of an area to neighboring communities and may be of help in understanding migration patterns. Family tree mapping using online mapping tools such as Google Earth (particularly when used with Historical Map overlays such as those from the David Rumsey Historical Map Collection) assist in the process of understanding the significance of geographical locations.

Dates

It is wise to exercise extreme caution with dates. Dates are more difficult to recall years after an event, and are more easily mistranscribed than other types of genealogical data. Therefore, one should determine whether the date was recorded at the time of the event or at a later date. Dates of birth in vital records or civil registrations and in church records at baptism are generally accurate because they were usually recorded near the time of the event. Family Bibles are often a source for dates, but can be written from memory long after the event. When the same ink and handwriting is used for all entries, the dates were probably written at the same time and therefore will be less reliable since the earlier dates were probably recorded well after the event. The publication date of the Bible also

provides a clue about when the dates were recorded since they could not have been recorded at any earlier date.

People sometimes reduce their age on marriage, and those under "full age" may increase their age in order to marry or to join the armed forces. Census returns are notoriously unreliable for ages or for assuming an approximate death date. Ages over 15 in the 1841 census in the UK are rounded down to the next lower multiple of five years.

Although baptismal dates are often used to approximate birth dates, some families waited years before baptizing children, and adult baptisms are the norm in some religions. Both birth and marriage dates may have been adjusted to cover for pre-wedding pregnancies.

Calendar changes must also be considered. In 1752, England and her American colonies changed from the Julian to the Gregorian calendar. In the same year, the date the new year began was changed. Prior to 1752 it was 25 March; this was changed to 1 January. Many other European countries had already made the calendar changes before England had, sometimes centuries earlier. By 1751 there was an 11 day discrepancy between the date in England and the date in other European countries.

The French Republican Calendar or French Revolutionary Calendar was a calendar proposed during the French Revolution, and used by the French government for about 12 years from late 1793 to 1805, and for 18 days in 1871 in Paris. Dates in official records at this time use the revolutionary calendar and need "translating" into the Gregorian calendar for calculating ages etc. There are various websites which do this.

Occupations

Occupational information may be important to understanding an ancestor’s life and for distinguishing two people with the same name. A person’s occupation may have been related to his or her social status, political interest, and migration pattern. Since skilled trades are often passed from father to son, occupation may also be indirect evidence of a family relationship.

It is important to remember that a person may change occupations, and that titles change over time as well. Some workers no longer fit for their primary trade often took less prestigious jobs later in life, while others moved upwards in prestige. Many unskilled ancestors had a variety of jobs depending on the season and local trade requirements. Census returns may contain some embellishment; e.g., from labourer to mason, or from journeyman to master craftsman. Names for old or unfamiliar local occupations may cause confusion if poorly legible. For example, an ostler (a keeper of horses) and a hostler (an innkeeper) could easily be confused for one another. Likewise, descriptions of such occupations may also be problematic. The perplexing description "ironer of rabbit burrows" may turn out to describe an ironer (profession) in the Bristol district named Rabbit Burrows. Several trades have regionally preferred terms. For example, "shoemaker" and "cordwainer" have the same meaning. Finally, many apparently obscure

jobs are part of a larger trade community, such as watchmaking, framework knitting or gunmaking.

Occupational data may be reported in occupational licenses, tax assessments, membership records of professional organizations, trade directories, census returns, and vital records (civil registration). Occupational dictionaries are available to explain many obscure and archaic trades.

Reliability of sources

Information found in historical or genealogical sources can be unreliable and it is good practice to evaluate all sources with a critical eye. Factors influencing the reliability of genealogical information include: the knowledge of the informant (or writer); the bias and mental state of the informant (or writer); the passage of time and the potential for copying and compiling errors.

The quality of census data has been of special interest to historians, who have investigated reliability issues.

Knowledge of the informant

The informant is the individual who provided the recorded information. Genealogists must carefully consider who provided the information and what he or she knew. In many cases the informant is identified in the record itself. For example, a death certificate usually has two informants: a physician who provides information about the time and cause of death and a family member who provides the birth date, names of parents, etc.

When the informant is not identified, one can sometimes deduce information about the identity of the person by careful examination of the source. One should first consider who was alive (and nearby) when the record was created. When the informant is also the person recording the information, the handwriting can be compared to other handwriting samples.

When a source does not provide clues about the informant, genealogists should treat the source with caution. These sources can be useful if they can be compared with independent sources. For example, a census record by itself cannot be given much weight because the informant is unknown. However, when censuses for several years concur on a piece of information that would not likely be guessed by a neighbor, it is likely that the information in these censuses was provided by a family member or other informed person. On the other hand, information in a single census cannot be confirmed by information in an undocumented compiled genealogy since the genealogy may have used the census record as its source and might therefore be dependent on the same misinformed individual.

Motivation of the informant

Even individuals who had knowledge of the fact, sometimes intentionally or unintentionally provided false or misleading information. A person may have lied in order to obtain a government benefit (such as a military pension), avoid taxation, or cover up an embarrassing situation (such as the existence of a non-marital child). A person with a distressed state of mind may not be able to accurately recall information. Many genealogical records were recorded at the time of a loved one's death, and so genealogists should consider the effect that grief may have had on the informant of these records.

The effect of time

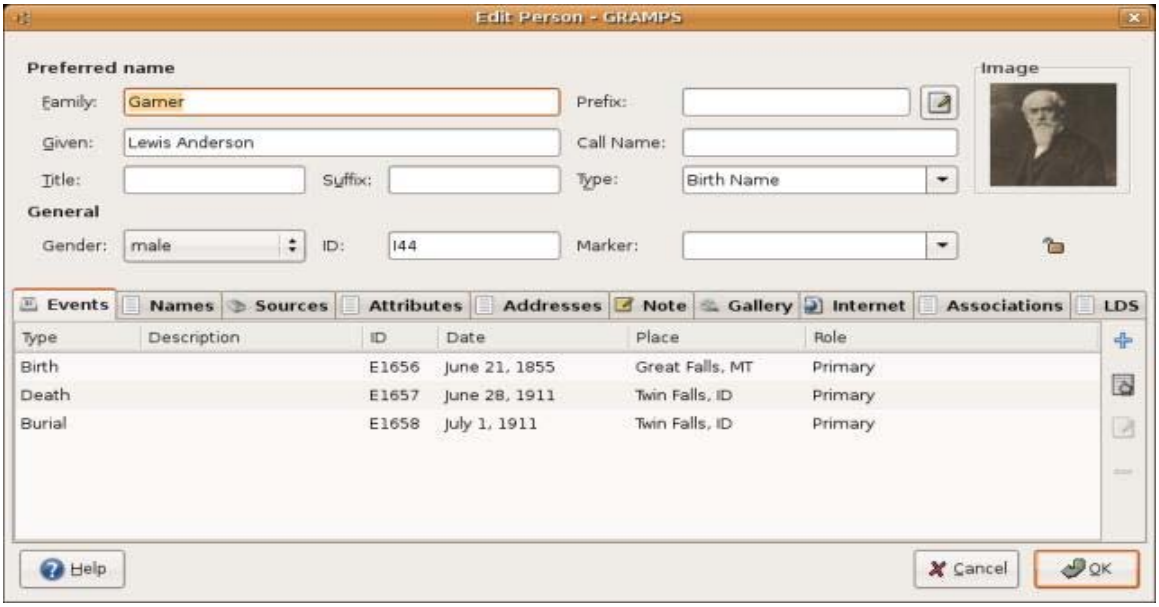
The passage of time often affects a person's ability to recall information. Therefore, as a general rule, data recorded soon after the event is usually more reliable than data recorded many years later. However, some types of data are more difficult to recall after many years than others. One type especially prone to recollection errors is dates. Also the ability to recall is affected by the significance that the event had to the individual. These values may have been affected by cultural or individual preferences.

Copying and compiling errors

Genealogists must consider the effects that copying and compiling errors may have had on the information in a source. For this reason, sources are generally categorized in two categories: original and derivative. An original source is one that is not based on another source. A derivative source is information taken from another source. This distinction is important because each time a source is copied, information about the record may be lost and errors may creep in from the copyist misreading, mistyping, or miswriting the information. Genealogists should consider the number of times information has been copied and the types of derivation a piece of information has undergone. The types of derivatives include: photocopies, transcriptions, abstracts, translations, extractions, and compilations.

In addition to copying errors, compiled sources (such as published genealogies and online pedigree databases) are susceptible to misidentification errors and incorrect conclusions based on circumstantial evidence. Identity errors usually occur when two or more individuals are assumed to be the same person. Circumstantial or indirect is evidence that does not explicitly answer a genealogical question, but either may be used with other sources to answer the question, suggest a probable answer, or eliminate certain possibilities. Compilers sometimes draw hasty conclusions from circumstantial evidence without sufficiently examining all available sources, without properly understanding the evidence, and without appropriately indicating the level of uncertainty.

Software



GRAMPS is an example of genealogy software

Genealogy software is computer software used to collect, store, sort, and display genealogical data. At a minimum, genealogy software accommodates basic information about individuals, including births, marriages, and deaths. Many programs allow for additional biographical information, including occupation, residence, and notes, and most also offer a method for keeping track of the sources for each piece of evidence.

Most programs can generate basic kinship charts and reports, allow for the import of digital photographs and the export of data in the GEDCOM format so that data can be shared with those using other genealogy software. More advanced features include the ability to restrict the information that is shared, usually by removing information about living people out of privacy concerns; the import of sound files; the generation of family history books, web pages and other publications; the ability to handle same sex marriages and children born out of wedlock; searching the Internet for data; and the provision of research guidance.

Programs may be geared toward a specific religion, with fields relevant to that religion, or to specific nationalities or ethnic groups, with source types relevant for those groups.

Confucius

The family tree of Confucius has been maintained for over 2,500 years, and is listed in the Guinness Book of Records as the largest extant family tree. The fifth edition of the Confucius Genealogy will be printed in 2009 by the Confucius Genealogy Compilation Committee (CGCC).