PROD #: V93N4A03

# The Science of Molecular Genealogy

*By* Ugo A. Perego; Ann Turner, M.D.;
Jayne E. Ekins; and Scott R. Woodward, Ph.D.

*Molecular science can help genealogists uncover previously unknown family relationships, verify or refute claims to ancestry, and shed light on questions that have puzzled family historians for years.*

All individuals carry a record of their ancestors in a complex chemical compound found inside almost every cell. Analysis of this molecule—deoxyribonucleic acid (DNA)—can help genealogists trace male- and female-line ancestors, prove and disprove relationships, reveal undocumented illegitimacies and adoptions, and identify familial ethnic and geographic origins.

DNA is packaged in threadlike structures called "chromosomes." Humans receive twenty-three chromosomes from each parent and, in turn, give half of their own DNA to each of their children. Parents, therefore, funnel a molecular record of their ancestors to their descendants.

More than 99 percent of each person's DNA is identical to that of all other people. This shared inheritance defines humans, yet the remaining 1 percent contains enough variation to make each person unique. The DNA of two closely related people has more similarities than that of distant cousins. Consequently, similarities and differences in DNA can show how closely individuals are related.

Molecular genealogists—also called "genetic genealogists"—test DNA samples from living individuals. Used in isolation, DNA test results have little value for family historians. Combined with documentary genealogical research, however, DNA evidence can help researchers identify ancestors and reconstruct family histories and lineages. Suppose, for example, that research reveals a candidate for a male ancestor's father but does not prove the relationship. If DNA samples from living male-line descendants of both men are different, they will disprove the hypothesis. If the samples match, however, the DNA alone does not

 ugo@smgf.org; Ann Turner, M.D., DNACousins@aol.com; Jayne E. Ekins, jayne@smgf.org; and Scott R. Woodward, Ph.D., scott@smgf.org. Mr. Perego is director of operations for the Sorenson Molecular Genealogy Foundation (SMGF), where he supervises the collecting of genetic and genealogical data for the foundation's worldwide database. In the past five years, he has written a dozen articles and given more than one hundred lectures on molecular genealogy. Dr. Turner founded the GENEALOGY-DNA mailing list at RootsWeb and co-authored (with Megan Smolenyak) *Trace Your Roots with DNA: Using Genetic Tests to Explore Your Family Tree* (Emmaus, Pa.: Rodale, 2004). Ms. Ekins, a research scientist with SMGF, coordinates genetic data production and analysis and performs original research. Dr. Woodward, whose work has been featured internationally, is chief scientific officer for SMGF and principal investigator for the molecular genealogy research project. His research interests include reconstructing ancient and modern genealogies using DNA techniques on samples worldwide, tracing human population movements by following gene migrations, and analyzing DNA found in ancient manuscripts.

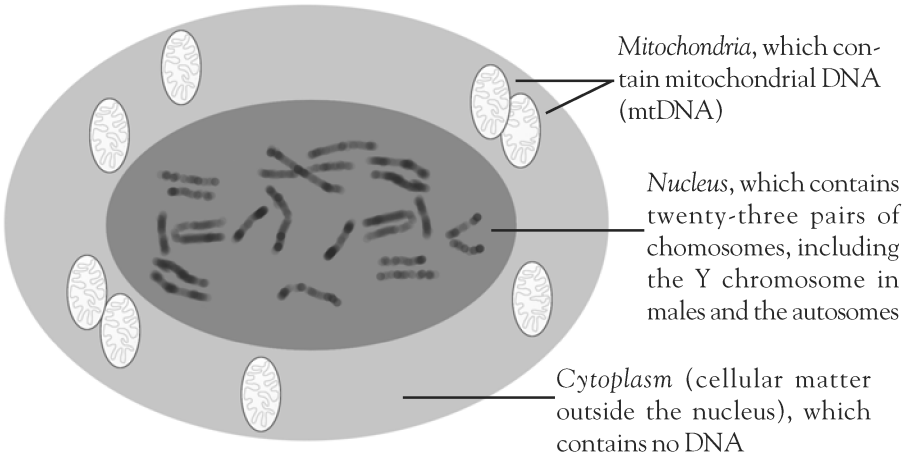
prove a father-son relationship but, in combination with documentary evidence, it could make the case persuasive.[1]

## GENEALOGICALLY USEFUL PATHWAYS OF GENETIC INHERITANCE

Each parent contributes approximately half of his or her child's DNA. Scientists usually cannot identify which parent provided which part of the child's DNA without testing one or both parents. The parts that researchers *can* identify, however, allow them to answer many genealogical questions: the Y chromosome, which is found in each cell's nucleus in males only, and mitochondrial DNA (mtDNA), which is found in each cell's cytoplasm. See figure 1.

Figure 1

**Selected Parts of a Human Cell**




### *The Paternal Lineage Pathway*

Sons receive a Y chromosome, usually unchanged, from their fathers. Occasionally, however, a slight alteration (called a "mutation") will occur in a random male's Y chromosome. A man with such an altered Y chromosome will pass it to his sons and they to their sons. Subsequently, all of their male-line descendants will pass that slightly altered Y chromosome to their sons. Further random mutations may occur occasionally in later generations. Thus, every living male's Y chromosome today carries a cumulative history of many small changes that have occurred in his paternal lineage over hundreds of generations and thousands of

1. Several recent books cover the basic biology of genetics and DNA as they apply to genealogical testing. See, for example, Chris Pomery, *DNA and Family History* (Toronto, Ont.: Dundurn, 2004); Thomas H. Shawker, *Unlocking Your Genetic History* (Nashville, Tenn.: Rutledge Hill, 2004); and Smolenyak and Turner, *Trace Your Roots with DNA*.

years. Because different changes occurred in different males over the millennia, their male descendants bear Y chromosomes with distinctive patterns, called "haplotypes," which can differentiate their families and ancestors.

The Y chromosome is useful for answering genealogical questions because it passes intact from generation to generation and its inheritance follows surnames in many western and some nonwestern societies. All male-line descendants of the same male ancestor—typically those with the same surname in these societies—will have the same or a very similar Y chromosome.[2] For example, residents of Tristan da Cunha, which has genealogical records dating from 1816, have eight Y-chromosome haplotypes corresponding to those of seven of the island's founders, whose surnames the residents bear, and an apparent visitor with an unknown surname.[3] In cases where the Y-chromosome haplotype did not correspond to the surname, it indicated ancestry more accurately than documentary genealogy or oral history.

Searching a database of Y-chromosome haplotypes paired with surnames can enable genealogists to identify relatives and disprove erroneous lineages. Males with the same surname and different haplotypes probably descend from different lines bearing the same surname. Conversely, similar haplotypes of males with different surnames might indicate adoption, illegitimacy, or other situations where names may have been altered somewhere in a male-line descent. For example, men with Lorentz and Lawrence surnames and the same Y-chromosome haplotype very likely descend from the same male-line ancestor.[4]

Two studies popularized applying Y-chromosome analysis to genealogical research, the highly publicized Jefferson-Hemings case and a study involving the Jewish priestly class of Cohen:

- The question of whether Thomas Jefferson fathered some or all of his slave Sally Hemings's children arose during his lifetime, and it is still the subject of debate today. Jefferson left no male issue through his wife, but living male-line descendants of his father's brother, who carried the same Y chromosome as Jefferson, were tested to determine the Jefferson haplotype. Also tested were male-line descendants of (1) Jefferson's brother-in-law John Carr (because of rumors that members of his family had fathered Hemings's children), (2) Sally Hemings's son Thomas Woodson, and (3) another Hemings son, Eston. Of the three lines, only the male-line descendants of Eston Hemings carry the Jefferson Y-chromosome haplotype.[5] The

2. Bryan Sykes and Catherine Irven, "Surnames and the Y Chromosome," *American Journal of Human Genetics* 66 (March 2000): 1417–19; and Mark A. Jobling, "In the Name of the Father: Surnames and Genetics," *TRENDS in Genetics* 17 (June 2001): 353–57.

3. Himla Soodyall and others, "Genealogy and Genes: Tracing the Founding Fathers of Tristan da Cunha," *European Journal of Human Genetics* 11 (September 2003): 705–9.

4. Ann Turner, "One or Many? Ann Turner Looks at the Role of DNA in the Study of Surname Origins," *Family Chronicle* 9 (March/April 2005): 46–49.

5. Eugene A. Foster and others, "Jefferson Fathered Slave's Last Child," *Nature* 396 (5 November 1998): 27–28.

DNA evidence alone does not prove that Jefferson was Eston's father, but it complements evidence drawn from other sources.[6]

- Researchers found that a noticeable fraction of Jewish priests share a common Y-chromosome haplotype, whether or not they are part of the far-flung Ashkenazi or Sephardic communities.[7] A later study found the same haplotype in the Lemba of southern Africa, a tribe with customs reminiscent of Jewish practices and an oral tradition that their ancestors came from the north by boat.[8] Finding the same haplotype in geographically dispersed groups implies descent from a single common ancestor.

Businessman Bennett Greenspan hoped that the approach used in the Jefferson and Cohen research would help family historians. After reaching a brick wall on his mother's surname, Nitz, he discovered an Argentine researching the same surname. Greenspan enlisted the help of a male Nitz cousin. A scientist involved in the original Cohen investigation tested the Argentine's and Greenspan's cousin's Y chromosomes. Their haplotypes matched perfectly. Furthermore, the haplotype did not match any of two dozen samples collected by Greenspan to serve as controls. Fortified by this demonstration that DNA could reflect a common lineage, Greenspan founded a private company offering DNA tests for genealogical purposes. His business was shortly followed by a half-dozen similar companies in the United States and Europe.[9]

More than two thousand Y-chromosome surname studies are underway, some with hundreds of participants.[10] Family historians interested in joining a project can find lists of active investigations on commercial testing company Web sites and Ancestry.com and Genforum.com message boards.[11] Many surname-project Web sites report genetic findings, and genealogical periodicals are beginning to carry case studies that include Y-chromosome analyses. Several examples demonstrate different genealogical uses of Y-chromosome data:

- Hundreds of men named Wells participated in Y-chromosome testing. Genealogical data collected prior to the project suggested twenty-four distinct families. The

6. Helen Leary, "Sally Hemings's Children: A Genealogical Analysis of the Evidence," and Thomas W. Jones, "The 'Scholars Commission' Report on the Jefferson-Hemings Matter: An Evaluation by Genealogical Standards," *NGS Quarterly* 89 (September 2001): 165–207 and 208–18.

7. Mark G. Thomas and others, "Origins of Old Testament Priests," *Nature* 394 (9 July 1998): 138–40.

8. Mark G. Thomas and others, "Y-chromosomes Traveling South: The Cohen Modal Haplotype and the Origins of the Lemba—the 'Black Jews of Southern Africa'," *American Journal of Human Genetics* 66 (February 2000): 674–86.

9. Bennett Greenspan, "An Insider's Look at the Genealogy DNA Field," *New England Ancestors* (Summer 2004): 21–23.

10. Bill Davenport, "Surname Projects: 'Over Fifty List'," *World Families Network* (http://worldfamilies.net/over50list.html).

11. The largest companies with family DNA projects are DNA Heritage (http://www.dnaheritage.com), Family Tree DNA (http://www.familytreedna.com), and Relative Genetics (http://www.relativegenetics.com). A more complete listing can be found at Megan Smolenyak Smolenyak, "Genetealogy Resources," *Genetealogy.com* (http://genetealogy.com).

Y-chromosome surname study, however, demonstrates that five presumed connections, based on similar names, dates, and places, are separate lines.[12]

- Y-chromosome samples from just two people solved a problem that had baffled researchers for years. Justin Howery and Fred Hauri, who believed that everyone with a variant of their surnames descends from a man who lived in the 1400s in the Swiss village of Beromuenster, could not document a family connection. Genetic testing, however, revealed that both men have the same Y-chromosome haplotype, even though their ancestors came to the United States from different countries in different centuries.[13]
- Although all Smolenyaks seem to trace their ancestry to one small village in Slovakia, they have four Y-chromosome haplotypes, indicating four distinct ancestral lines.[14]
- The ancestral haplotype of Edmund Rice, who immigrated to Massachusetts in 1638, was established by matching DNA results from descendants of five different sons. The testing also revealed a "non-paternity event"—possibly an unrecorded adoption or illegitimacy—in one line of male descent.[15]

*The Maternal Lineage Pathway*

In addition to the DNA in the nucleus of most cells, DNA is also found in structures called mitochondria in each cell's cytoplasm. See figure 1. Cells have hundreds of mitochondria, each containing many DNA molecules called "mitochondrial DNA" or "mtDNA." The mother's—but not the father's—mitochondria are present in the fertilized egg that is the first cell of a new human being. Thus, a mother passes her mtDNA to her sons and daughters. Her daughters, but not her sons, pass their mtDNA to the next generation. Consequently, mtDNA, inherited exclusively from mothers, passes intact from generation to generation.

Just as with the Y chromosome, slight random changes in mtDNA molecules over many generations result in different patterns or haplotypes. However, these mutations occur less frequently than in Y chromosomes. The mutation rate has been measured in Iceland, which has genealogical records covering many generations. Only three mutations occurred in 705 "transmission events" (opportunities for a mutation to occur between generations). Some of the residents with matching mtDNA haplotypes were twelve generations removed from their common female-line ancestor, who was born in 1560.[16]

---

12. Ken Wells, "Relative Advance: DNA Testing Helps Find Family Roots," *Wall Street Journal*, 6 March, 2003, page A1.

13. Justin Howery, "Howery DNA Project," message board posting, 17 November 2000, *GENEALOGY-DNA-L Archives* (http://archiver.rootsweb.com/th/read/GENEALOGY-DNA/2000–11/0974503831).

14. Megan Smolenyak Smolenyak, "DNA Testing Dispels a Genealogical Myth," *Everton's Family History Magazine* 56 (May/June 2002): 44–48.

15. Robert V. Rice and John F. Chandler, "DNA Analyses of Y-chromosomes Show Only One of Three Sons of Gershom Rice to be a Descendant of Edmund Rice," *New England Ancestors* 3 (Fall 2002): 50–51.

16. Sigrun Sigurgardottir and others, "The Mutation Rate in the Human mtDNA Control Region," *American Journal of Human Genetics* 66 (May 2000): 1599–1609.

The popularity of mtDNA for genealogical purposes followed the use of mtDNA to confirm the identity of remains thought to be those of the wife and children of Nicholas II, czar of Russia. The mtDNA extracted from the remains matched that of living relatives who shared a common maternal line with the czar's wife.[17] Such testing can disprove relationships as well: the mtDNA of Anna Anderson Manahan, who claimed to be Nicholas's daughter Anastasia, did not match that of the czar's family.[18]

*Ethnic and Geographic Pathways*

"Autosomes" are twenty-two pairs of chromosomes that children inherit from both parents, and the DNA they contain is called "autosomal DNA." Autosomes do not include the sex chromosomes (X and Y) and mtDNA. In contrast to mtDNA and Y-chromosome molecules, which parents pass intact to their children, autosomal DNA comprises a random combination of both parents' genetic makeup. Because each parent's autosomal DNA recombines, the half that each transmits to a child is a mixture of the DNA that the contributing parents received from their parents. Consequently, if neither parent's autosomal DNA is studied, it is not possible to determine which parent or ancestor contributed any segment of the child's autosomal DNA. In addition, the proportion of ancestral genetic contribution to a descendant decreases with the number of generations between the ancestor and descendant. Both the recombination of autosomal DNA and its decreasing proportions over generations create challenges for using autosomal DNA for genealogical purposes.

Autosomal DNA is the focus of techniques to determine origins more generally than paternal and maternal lines allow. Genealogists can submit DNA samples to a company that tests autosomal DNA to identify continental or subcontinental origins.[19] In a broad sense, individuals can use the results to reconnect with family roots that may have been previously unknown to them.

Other research has focused on inferring ancestry more specific than broad geographic or ethnic classifications. A recently proposed approach allows a participant's assignment to a hierarchical set of populations. For example, the first-level test results may imply European origins. Subsequent levels may narrow the inference successively to the British Isles, a region in southwest Wales, and perhaps an extended family from the area. Such inference of ancestry from all

17. Peter Gill and others, "Identification of the Remains of the Romanov Family by DNA Analysis," *Nature Genetics* 6 (February 1994): 130–35. Recent publications have challenged the findings on the basis of difficulties involved in recovering and analyzing ancient DNA. See Alex Knight and others, "Molecular, Forensic and Haplotypic Inconsistencies Regarding the Identity of the Ekaterinburg Remains," *Annals of Human Biology* 31 (March–April 2004): 129–38.

18. Peter Gill and others, "Establishing the Identity of Anna Anderson Manahan," *Nature Genetics* 9 (January 1995): 9–10.

19. Tony Frudakis and others, "A Classifier for SNP-Based Racial Inference," *Journal of Forensic Science* 48 (July 2003): 771–82. For further information, see Tony N. Frudakis, "Powerful but Requiring Caution: Genetic Tests of Ancestral Origins," in the present issue of the *NGS Quarterly*.

areas of the world is possible, but accuracy depends on the depth of sampling from each region. Participants receive "likelihood scores" for each level, which enable them to weigh the results. Assignment to broad ethnic and geographic classifications applies to questions of deeply rooted ancestry. In contrast, inferring membership in more localized populations can provide information on a genealogically useful scale.[20]

## DNA IN THE LABORATORY

Today commercial laboratories apply techniques that are spin-offs from the Human Genome Project, the massive international collaboration to analyze the complete set of human chromosomes.[21] Nevertheless, current technology has not yet advanced to the point where laboratories can report an individual's entire genetic makeup.[22] Instead, specialized tests analyze limited sections of DNA to help solve problems in areas including crime investigation, paternity, identification of human remains, and genealogy.

### *Mitochondrial DNA Sequences*

DNA consists of long sequences of four chemical compounds. These building blocks—called "bases" or "nucleotides"—are often abbreviated as A (adenine), C (cytosine), G (guanine), and T (thymine). Scientists were initially skeptical that such a limited set of chemicals could account for the complexity of life. However, the four bases can be arranged in many different orders, just as letters from the English alphabet can be shuffled in many different combinations and chained together into a complete book. The human genome contains about three billion bases. Some of its sections can be decoded, but a surprisingly large amount, perhaps as much as 98 percent, appears to be meaningless. Genealogical tests focus on these regions of "junk" DNA, and thus they cannot reveal any personal traits or medical conditions. For an example of a short sequence of DNA bases, see table 1.

A cell's mtDNA contains 16,569 sequenced bases but, for genealogical questions, laboratories typically study segments containing only 400 to 1,100 of the most informative bases. These sections are called "hypervariable" because they show more differences among people than mtDNA's other regions. Because a report listing even 400 bases would be difficult to interpret, laboratories conducting mtDNA tests customarily report only the bases that differ from a standard sequence called the Cambridge Reference Sequence (CRS). For example, a re-

20. Jayne E. Ekins and others, "Inference of Ancestry: Constructing Hierarchical Reference Populations and Assigning Unknown Individuals," *Human Genomics* (forthcoming).

21. "The Human Genome Project Completion: Frequently Asked Questions," *National Human Genome Research Institute* (http://www.genome.gov/11006943).

22. Corie Lok, in "Deciphering DNA, Top Speed," *TechnologyReview.com* (http://www.technologyreview.com/articles/05/05/issue/forward_dna.asp?p=1), writes "Using about 100 state-of-the-art sequencing machines to fully sequence the 3.2 billion DNA letters that make up one person's genome would take six months and cost $20 million to $30 million."

Table 1

**Example of a Single Nucleotide Polymorphism (SNP)**

*Reference sequence in a DNA segment:*

TAGCCGATACGACGATACGGACTAGACGTACACCCATCGTTACG

*Comparison, showing a SNP in the seventh position in the segment:*

TAGCCG**T**TACGACGATACGGACTAGACGTACACCCATCGTTACG

*Note:* The letters in the DNA segments refer to the four chemical compounds (called "bases" or "nucleotides") that comprise the DNA molecule: adenine (A), cytosine (C), guanine (G), and thymine (T).

port of an "HVR 1" test result as "16093C" would mean that the subject's mtDNA in hypervariable region one (HVR 1)—an mtDNA segment containing bases in positions numbered 16,024 through 16,365—differs from the CRS because it has a cytosine (C) base at position 16,093. Most people have a few differences from the CRS.

*The Smallest Changes in DNA*

As described above, mutations are modifications in DNA molecules that occur randomly. Mutations can have positive or negative effects, but they typically occur in sections of DNA that have no effect. A mutation that replaces just one base (nucleotide) with another is called a "single nucleotide polymorphism" or "SNP" (pronounced "snip"). For an example, see table 1.

The changes in the mtDNA molecule described above are literally SNPs, but more often the term is applied to SNPs sprinkled throughout the chromosomes, including the Y. SNPs, which tend to be rare, often represent unique events. Given their low rate, SNPs are used in anthropological studies for tracing extremely deeply rooted pedigrees, for example, determining matrilineal or patrilineal descent from one of several ancient "clans."[23]

By using multiple SNPs researchers can determine the order in which the SNPs occurred and estimate when two ancient lineages diverged. See table 2. The variability of SNPs among descendants of a "founding father" gives a rough estimate of when he lived: the more SNPs the descendants have, the more time has elapsed since their lineages diverged. For example, The Y Chromosome Consortium has identified a set of SNPs useful in classifying males into hierarchically related clusters. Small clusters with identical haplotypes can be combined into larger groups with similar but not identical haplotypes and so forth

23. Bryan Sykes, *The Seven Daughters of Eve: The Science that Reveals our Genetic Ancestry* (New York: W. W. Norton, 2001).

Table 2

**Single Nucleotide Polymorphisms (SNPs)**
**Occurring at Different Times in a DNA Segment**

*Reference Sequence*

TAGCCGATACGACGATACGGACTAGACGTACACCCATCGTTACG

*Comparison Sequence 1*

TAGCCG**T**TACGACGATACGGACTAGACGTACACCCATCGTTACG

*Comparison Sequence 2*

TAGCCG**T**TACGACGATACGGACTAGACGTACACCCA**A**CGTTACG

*Comparison Sequence 3*

TAGCCG**T**TACGACGATACGGACTA**C**ACGTACACCCATCGTTACG

*Note:* The letters in the DNA segments refer to the four chemical compounds (called "bases" or "nucleotides") that comprise the DNA molecule: adenine (A), cytosine (C), guanine (G), and thymine (T).

The mutation from an "A" to a "T" at position 7 appears in all three comparison sequences. Sequences 2 and 3 each have an additional mutation, so Sequence 1 is the oldest of the three comparison sequences. It is not possible to tell whether Sequence 2 or 3 is the next oldest.

until the groups are joined into a tree representing all humankind. The clusters, labeled in a systematic alphanumeric fashion similar to an outline, comprise haplogroups and subhaplogroups.[24] Because different haplogroups predominate in different regions of the world, Y-chromosome and mtDNA haplogroups suggest ethnic and geographic origins of patrilineal and matrilineal ancestry—for example a man in Y-chromosome group R1b might have male-line ethnic origins in Western Europe.[25]

*The Most Genealogically Useful Changes in DNA*

The genetic information that genealogists most often employ is the "short tandem repeat" (STR). The term refers to the repetition of a short sequence of bases. For instance, a sequence of four bases, like G-A-T-A, might occur seven consecutive times in one segment of DNA. When an STR mutates, the number of repetitions changes—for example seven repetitions of G-A-T-A at

24. "A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups," *The Y Chromosome Consortium* (http://ycc.biosci.arizona.edu/nomenclature_system/frontpage.html).

25. J. Douglas McDonald, "World Haplogroup Maps," *McDonald Group* (http://www.scs.uiuc.edu/~mcdonald/WorldHaplogroupsMaps.pdf).

one location on a father's Y chromosome might change to six repetitions of the same sequence at the same location on his son's Y chromosome. For an example see table 3. Such mutations pass unchanged from parent to child until another mutation occurs.

Table 3

**Examples of Short Tandem Repeats (STRs)**

*Sample 1: Seven repeats of the short sequence of DNA bases GATA*

TACCATGC … GATAGATAGATAGATAGATAGATAGATA … CTCGGT

*Sample 2: Six repeats of the GATA sequence*

TACCATGC … GATAGATAGATAGATAGATAGATA … CTCGGT

*Note:* The letters in the DNA segments refer to the four chemical compounds (called "bases" or "nucleotides") that comprise the DNA molecule: adenine (A), cytosine (C), guanine (G), and thymine (T).

As various distinct STR mutations accumulated in different lineages over many centuries, each developed its own pattern of STRs. Consequently, people with different lineages have distinct inherited STR patterns. Individuals with identical patterns are said to bear the same "haplotype," such as the Jefferson Y-chromosome haplotype of Eston Hemings's male-line descendants. Minor differences in haplotypes are compatible with descent from a common ancestor.

Genetic tests determine the pattern of STRs on the Y chromosome. Locations or segments on the chromosome often have labels starting with the letters DYS (for example, "DYS439"), which stand for "DNA Y-chromosome sequence."[26] Genetics laboratories test between twelve and forty locations on the Y chromosome. Their reports list the DYS numbers of the locations tested and the number of STRs in each location—such as 13 at DYS 393. For example, the DNA test that determined the Jefferson haplotype reported the number of STRs at eleven DYS locations on a Jefferson descendant's Y chromosome. See table 4.

## DNA ON THE INTERNET

### *Y-chromosome Databases*

Most Y-chromosome tests take place on a small scale within surname projects, such as the Edmund Rice study described above. However, large assemblies of Y-chromosome data are available on the Internet, which can place an individual's test results in a global context. Such public databases might generate privacy

26. John M. Butler, *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (Burlington, Mass.: Elsevier Academic Press, 2005), 23–25.

Table 4

**Three Unrelated Y-chromosome Haplotypes**

| | *Y-chromosome location labels* | | | | | | |
|---|---|---|---|---|---|---|---|
| | DYS 19 | DYS 389-1 | DYS 389-2 | DYS 390 | DYS 391 | DYS 392 | DYS 393 |
| | *Number of short tandem repeats (STRs) at each location* | | | | | | |
| Jefferson haplotype | 15 | 12 | 27 | 24 | 10 | 15 | 13 |
| Carr haplotype | 14 | 13 | 29 | 24 | 10 | 13 | 13 |
| Woodson haplotype | 17 | 14 | 31 | 21 | 10 | 11 | 14 |

*Source:* Megan Smolenyak and Ann Turner, *Trace Your Roots with DNA: Using Genetic Tests to Explore Your Family Tree* (Emmaus, Pa.: Rodale, 2004), 193. The table, using modern nomenclature and standards, converts eleven DYS locations reported in the original study to seven. For the original data, see Eugene A. Foster and others, "Jefferson Fathered Slave's Last Child," *Nature* 396 (5 November 1998): 27–28.

concerns, but they do not reveal personal identities, and their data do not contain information about personal traits or medical conditions. The rapidly accumulating volume of online genetic information can aid investigations of genealogical questions.

Results from the 1998 study of Carr, Jefferson, and Woodson haplotypes illustrate the useful information that can be gleaned from online databases. A centerpiece argument was the rarity of the Jefferson haplotype, suggesting that the Jefferson-Hemings match was not coincidental. When the article was written, the database of Y-chromosome haplotypes contained 670 European records, with no matches to the Jefferson haplotype.[27] Today, however, such databases have vastly larger numbers of records:

- Y-chromosome Haplotype Reference Database (YHRD) is an anonymous database of records submitted by forensic laboratories and collected to provide a cross-sample of people in specific locations. It contains 28,650 world-wide records, including 18,711 from Europe.[28]
- Ybase was the first publicly accessible database that allowed individuals who had used different testing companies to enter their data and compare results. It contains 5,025 Y-chromosome haplotypes, 6,214 surnames, and useful statistical summaries showing the range of STRs for tested Y-chromosome locations.[29]

27. Foster and others, "Jefferson Fathered Slave's Last Child," 27–28.

28. "About the 'YHRD - Y Chromosome Haplotype Reference Database'," *YHRD.org* (http://www.yhrd.org).

29. *Ybase: Genealogy by Numbers* (http://www.ybase.org).

- Ysearch is a publicly accessible database containing approximately thirteen thousand Y-chromosome records with the large majority having test results for twelve to twenty-four markers. Anyone can manually add data obtained from any company, and an automated procedure is available for Family Tree DNA customers. Users can add genealogical information to their Y-chromosome records.[30]
- The Sorenson Molecular Genealogy Foundation (SMGF) database includes 13,489 Y chromosomes linked to 550,000 ancestors.[31] With 9,400 unique surnames and more than 90 percent of the Y-chromosome haplotypes tested at thirty or more markers, it is the largest searchable Y-chromosome database. SMGF analyzes samples contributed by volunteers, who can order a free participation kit from the Web site.[32]
- Ymatch contains thirty-five hundred records with the majority having test results for twenty-six to forty-three Y-chromosome locations. This is the newest database of this kind available to the public.[33]

None of the above sources contains a match for the Jefferson haplotype. In contrast, the Sorenson database has 531 records, with many different surnames, matching the Carr haplotype shown in table 4, which was based on eleven markers (seven markers, when applying modern standards). Testing more markers on the Carr sample, as various companies do today, would reduce the matches to those most closely related to the Carrs who were tested.

The results for the Woodson haplotype, depicted in table 4, are instructive. The Sorenson database has just one match, a sample from Ghana. That does not necessarily mean that the Woodson line came from Ghana—a larger database could show matches in other localities. However, this result suggests that the haplotype is more typical of African ancestry than European. Such geographical information may shed light on ancestral lines that lack documentary evidence for their origins.

*Mitochondrial DNA Databases*

Just as with Y-chromosome analysis, a person with mtDNA test results can compare them with various online databases. DNA data for the "Ice Man" demonstrate the information available. A body discovered in 1991 at the edge of a melting Alpine glacier was initially thought to be a climber who had died in modern times. Scientists soon determined, however, that the remains were some five thousand years old.[34] The Ice Man had a relatively common mtDNA haplotype present in 1 to 2 percent of Europeans. Containing two differences from

30. *Ysearch* (http://www.ysearch.org).

31. Ugo A. Perego, Natalie M. Myres, and Scott R. Woodward, "'Y' Research Through DNA," *Everton's Family History Magazine* 58 (May–June 2004): 26–28.

32. *Sorenson Molecular Genealogy Foundation* (http://www.smgf.org).

33. *Relative Genetics* (http://www.relativegenetics.com).

34. Oliva Handt and others, "Molecular genetic analyses of the Tyrolean Ice Man," *Science* 264 (17 June 1994): 1775–78.

the Cambridge Reference Sequence (CRS), his sample had cytosine (C) bases at positions 16,224 and 16,311 in hypervariable region 1.[35]

The following online databases contain mtDNA test results:

- MitoMap lists positions where differences from the CRS have been reported. These are listed by mtDNA location number rather than as a composite haplotype. However, it can be seen that many studies have reported differences. Occasional individuals have a novel difference, one never previously described.[36]
- The Mitochondrial DNA Concordance is a collection of several thousand mtDNA haplotypes reported in the technical literature up to about 1998.[37] The Ice Man's mtDNA haplotype (16224[C] 16311[C]) can be found in two places, under the listings for locations 16,224 or 16,311. The database shows that the Ice Man's haplotype occurs in many European populations—some Basque, Bavarian, Bulgarian, Cornish, English, Finnish, German, Norwegian, Portuguese, Swiss, Tuscan, and Welsh people have the same haplotype. Thus it is not possible to ascribe the Ice Man's ancestral line to a specific European location; however, because the haplotype is not found on other continents, the broad classification of European ancestry is confirmed.
- Oxford Ancestors offers guest access to its database. Entering the Ice Man's mtDNA test results shows 250 matches. Oxford Ancestors classifies him as a member of the "Katrine" clan, a pseudonym for mitochondrial haplogroup K.[38]
- Mitosearch is a public-access database of individually contributed GEDCOM files and mtDNA test results that have not been independently verified. A recent survey of the database produced fifty-eight matches for the Ice Man. It also yielded 150 members of haplogroup K. They had the Ice Man's two differences from the CRS plus various additions, demonstrating variation within a haplogroup.[39]
- The mtDNA Log functions like a guest book. Contributors may leave free-form comments along with their mtDNA test results. The site does not have a search function, but the Web browser's "Find" function substitutes. Visitors often provide data about their ancestral names and geographical locations.[40]
- The Federal Bureau of Investigation maintains an "mtDNA Population Database," which incorporates sequences from the Mitochondrial DNA Concordance (above) as well as more recent contributions from accredited forensic testing laboratories. This anonymous database can reveal whether a haplotype is common or rare.[41]

35. Michael D. Coble and others, "Single Nucleotide Polymorphisms Over the Entire mtDNA Genome that Increase the Power of Forensic Testing in Caucasians," *International Journal of Legal Medicine* 118 (June 2004): 137–46.

36. *MITOMAP: MtDNA Control Region Sequence Polymorphisms* (http://www.mitomap.org/cgi-bin/mitomap/tbl6gen.pl : dated 27 September 2005).

37. Kevin Miller and John Dawson, *Mitochondrial DNA Concordance* (http://www.bioanth.cam.ac.uk/mtDNA/).

38. "Oxford Ancestors' Databases," *Oxford Ancestors* (http://oxfordancestors.com/members).

39. *Mitosearch* (http://www.mitosearch.org).

40. Charles F. Kerchner Jr., *Mitochondria DNA (mtDNA) Test Results Log (BLOG)* (http://www.kerchner.com/cgi-kerchner/mtdna.cgi).

41. Keith L. Monson and others, "The mtDNA Population Database: An Integrated Software and Database Resource for Forensic Comparison," *Forensic Science Communications* 4 (April 2002), electronic edition (http://www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm).

- GenBank is a repository at the National Institutes of Health for raw mtDNA sequences from technical literature. With some effort public users can align their sequences to the published data and compare the results.[42]
- Many individuals have developed custom Web sites devoted to a specific haplogroup. The World Families Network maintains a page of links to these special interest groups.[43] For instance, John Walden's Web site for haplogroup K diagrams the relationships between the "clan mother" for haplogroup K and the variations he has located. [44]

## DNA TESTING IN THE FUTURE

In years to come DNA tests probably will be faster and cheaper and will include more markers than today's tests. Publicly accessible databases of compiled genetic information will also continue to grow, allowing genealogists to correlate DNA test results with population-based studies, such as the National Geographic Society's recently launched Genographic Project.[45] Large databases containing mtDNA and Y-chromosome matches could suggest research pathways that might unblock a lineage problem that seems unsolvable with documentary research alone.

The bulk of current tests for genealogical purposes is limited to the Y chromosome and mtDNA. This is a severe constraint because straight paternal and maternal lineages represent only a tiny fraction of anyone's total ancestry and DNA. The other parts of the pedigree harbor vast amounts of information that future genetic testing might unlock. Laboratories are beginning to study the use of autosomal DNA for genealogical purposes.

## CONCLUSION

Molecular genealogy synthesizes traditional genealogical research and relatively new technologies developed to explore genetic characteristics of the world's people. The combination enhances traditional genealogical methods, especially when ambiguities and roadblocks in written records impede documentary research. Scientific methods are just beginning to tap into the invaluable repository of ancestral information that is carried in every individual's DNA. Molecular methods can help individuals uncover previously unknown family relationships, verify or refute claims to ancestry, and shed light on questions that have puzzled genealogists for years.

Currently the two most active areas of genetic testing for genealogical purposes focus on mtDNA and the Y chromosome. DNA projects for family history

42. "GenBank Overview," *National Center for Biotechnology Information* (http://www.ncbi.nlm.nih.gov/Genbank/).

43. "mtDNA—The Family of Woman," *World Families Network* (http://worldfamilies.net/mtDNA.htm).

44. John S. Walden, "Swinging in the mtDNA Tree," *Walden/Adams/Walts/Waltz Surname DNA Projects* (http://freepages.genealogy.rootsweb.com/~jswdna/mtdna.html).

45. NationalGeographic.com, *The Genographic Project* (https://www3.nationalgeographic.com/genographic).

purposes can use samples from only two participants or hundreds. Many questions can be approached by querying online searchable genetic databases to find genetic matches to a known DNA profile. These remarkable resources are freely available and continuously expanding. Molecular genealogy methods eventually will enable genealogists to explore lines beyond strictly matrilineal and patrilineal ancestry. In the near future genealogists can expect a burgeoning expansion of this field. Genetic testing will be more widely available, increasingly economical for the individual, and more informative for answering a greater variety of genealogical questions.

---

**Emigrated Before Birth?**

[William Spreen declaration of intent, Taylor County, Wisconsin, Declarations of Intention 4: 408, Wisconsin Historical Society, Madison; microfilm 2,134,516, Family History Library, Salt Lake City, Utah. Underlined portions are handwritten on a printed form.]

I, William Spreen, aged 28 years, occupation Farmer, do declare on oath that my personal description is: Color White, complexion Light, height 5 feet 7 inches, weight 145 pounds, color of hair Brown, color of eyes Blue[,] other visible distinctive marks None; I was born in New Stetten, Germany, on the 19 day of January, anno Domini 1890; I now reside at Medford Wisconsin. I emigrated to the United States of America from Old Stetten, Germany on the vessel Unknown; my last foreign residence was New Stetten, Germany. It is my bona fide intention to renounce forever all allegiance and fidelity to any foreign prince, potentate, state, or sovereignty, and particularly to William II German Emperor, of which I am now a subject; I arrived at the port of New York, in the State of New York on or about the 15 day of June, anno Domini 1889; I am not an anarchist; I am not a polygamist nor a believer in the practice of polygamy; and it is my intention in good faith to become a citizen of the United States of America and to permanently reside therein: So help me God. William Spreen (original signature of declarant.) Subscribed and sworn to before me this 8th day of January, anno Domini 1919. (seal) S. A. M$^{c}$Comber, Clerk of the Circuit Court.

—*Contributed by Joy Reisinger,* CG