World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

June 9, 2005

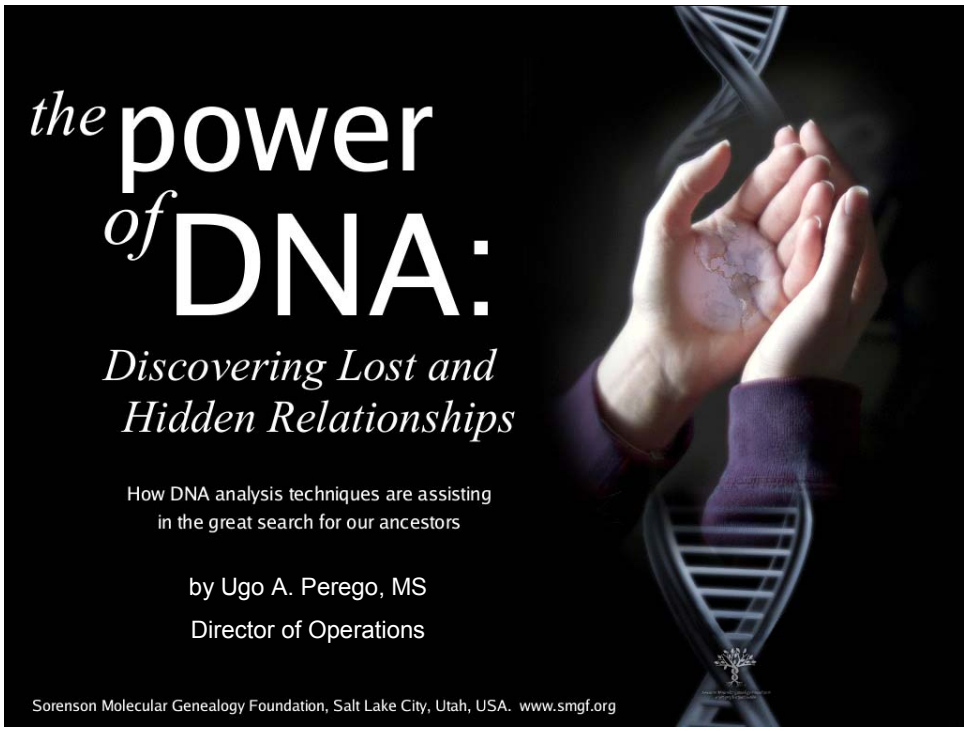
Code Number: 087-E
Meeting: 107 Genealogy and Local History with Geography and Map Libraries *(Part II)*

The Power of DNA: Discovering Lost and Hidden Relationships

How DNA Analysis techniques are assisting in the great search for our ancestors.

Ugo A. Perego, MS Director of Operations
Sorenson Molecular Genealogy Foundation
Salt Lake City, Utah
USA

Abstract:
The Sorenson Molecular Genealogy Foundation is building the world's largest database of correlated genetic and genealogical information to enable genealogical research to be performed using DNA analysis techniques. DNA samples with associated 4-generation pedigree charts have so far been collected from approximately 50,000 volunteers. Up to 170 regions of DNA are currently analyzed for each individual, and the corresponding pedigree chart is extended as far as genealogical databases allow, to currently include over 1,000,000 ancestral records. By combining these two sets of correlated data on an unprecedented scale, we are enabling progress for the first time into the new field of "molecular genealogy."
Molecular genealogy is the application of DNA analysis techniques and statistical population genetics to the task of reconstructing unknown genealogies from the genetic and genealogical information of living individuals. We address aspects of using DNA for genealogical research, including those of identification and differentiation of populations (with population boundaries defined not just by factors of demographic separation, but also by time periods), differences in inheritance models of the various types of genetic data, clustering, statistical reconstruction of ancestral trees, inference of ancestral genetic signatures, and inference of surname based on paternal-line DNA.



As the level of interest in genealogical research rises in the United States and abroad it becomes increasingly evident that written historical records, which are predominant sources for genealogical information, pose significant limitations to a genealogist's success in constructing a family tree. Some genealogists are exploring the possibility of augmenting traditional genealogical research methods with genetic testing.

Reference:
Maritz Marketing Research, Inc., *Sixty Percent of Americans Intrigued by Their Family Roots*, 2000 [online]. (http://www.maritzresearch.com/release.asp?rc=195&p=2&T=P). Accessed April 2005.

Goals of Molecular Genealogy

- To create a comprehensive database of the peoples of the world, using correlated genealogical and genetic information
- To provide the tools needed to reconstruct genealogies using DNA
- To change the way that we think about each other, and hopefully the way we act towards each other, by showing that we are really one great human family

SORENSON MOLECULAR GENEALOGY FOUNDATION
WWW.SMGF.ORG

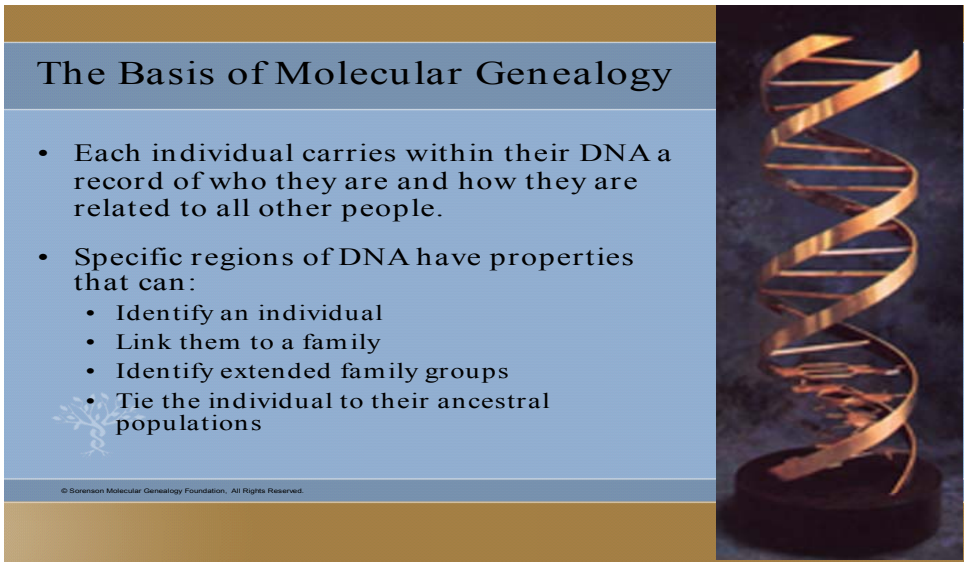
Genetic testing for genealogical purposes is a relatively new discipline with many techniques still in early stages of development. Although the technical aspects of these studies can be difficult to understand, the principles are familiar to most people.

The DNA Paradox

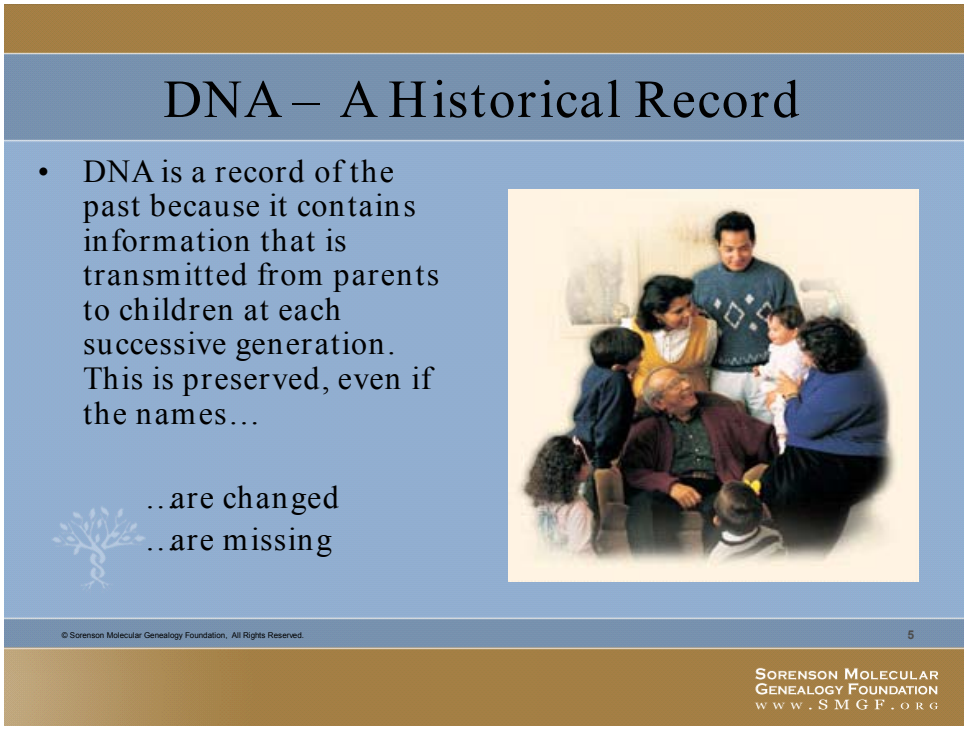
- Almost 4 billion pieces of information
- Can identify you as a unique individual
- All humans share many regions exactly
- The level of sharing is directly related to the degree of relationship
- DNA is what makes us *different*
- DNA is what makes us the *same*

SORENSON MOLECULAR GENEALOGY FOUNDATION
WWW.SMGF.ORG

Establishing family relationships using genetic analysis is based on the well-known phenomenon of inheritance. Inheritance occurs because every mother and father pass to each child deoxyribonucleic acid (DNA) in the form of chromosomes. The inheritance of DNA thus provides an unbreakable biological link between past, present and future generations.



Everyone has DNA. Almost every cell in the human body contains a copy of the genetic blueprint of that individual's life. Inside the nucleus of each cell there are 46 chromosomes, 23 inherited from the father and 23 inherited from the mother. In addition to nuclear DNA (the DNA found inside the nucleus), DNA is also found in small energy-producing organelle known as mitochondria. This DNA is therefore called mitochondrial DNA (mtDNA).
Specific regions (areas) of DNA can be studied to determine close and far relationships with other people, both in the present and in the past.



We are not only the recipients of DNA from the past, as we also give half of our unique genetic combination to each of our children, thus contributing to the genetic heritage that is passed onto the next generation. Every person is therefore an unbreakable link between forebears and descendants, where genetic information is funneled through us from the past, into the future. We are literally a walking, breathing, living record of our own family history.

Number of possible ancestors

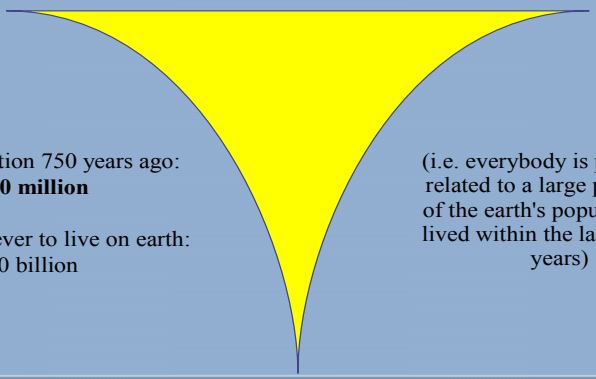
Generations	Ancestors	Years
1	2	25
2	4	50
3	8	75
4	16	100
10	1024	250
15	32768	375
20	1,048,567	500
25	33,554,432	625
30	1 billion	750
35	30 billion	875
40	1 trillion	1000

SORENSON MOLECULAR GENEALOGY FOUNDATION
WWW.SMGF.ORG

There is a simple mathematical rule to find out how many ancestors we had at EACH given generations. The formula is 2^n where n is the number of generations. For example 2^{10} would reveal how many ancestors we had 10 generations ago: 2x2x2x2x2x2x2x2x2x2 = 1,024 ancestors living at the 10th generation ONLY. These numbers add up quickly and only 30 generations ago, each one of us would have had 1,034,000,000 ancestors (this is over 1 billion). Now, if we estimate each generation to be approximately 25 years, it would means that 750 years ago (approximately the year 1250AD), each one of us would have had that many ancestors.

Problems with the numbers

30 generations = 750 years = **1 billion** possible ancestors



World population 750 years ago:
~ **450 million**

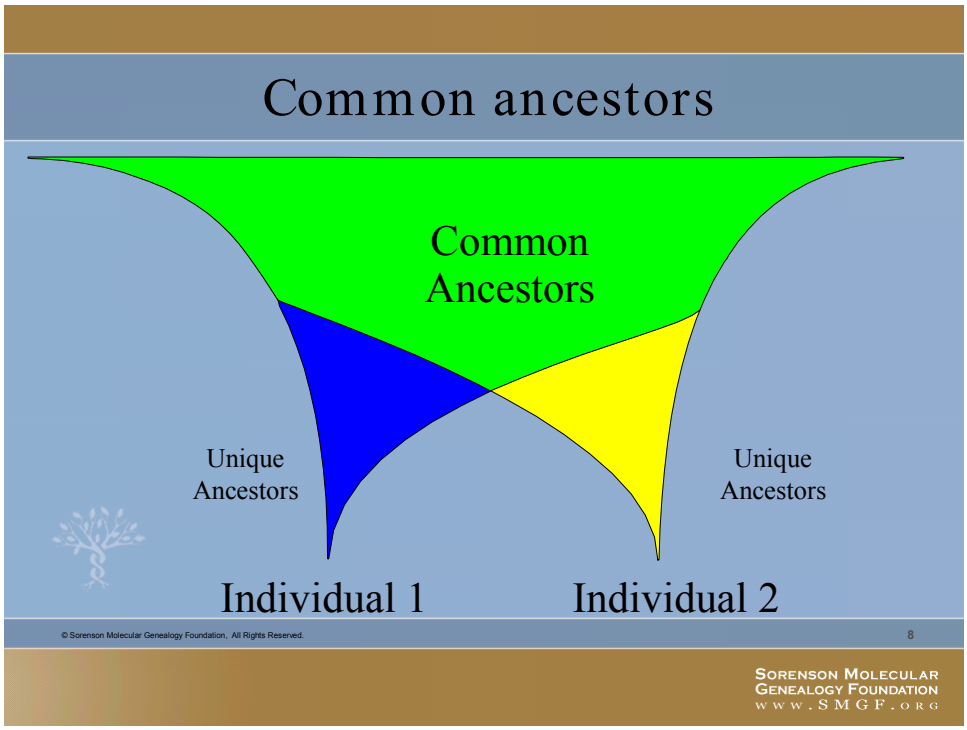
Total humans ever to live on earth:
~ 70 billion

(i.e. everybody is potentially related to a large proportion of the earth's population that lived within the last 500-750 years)

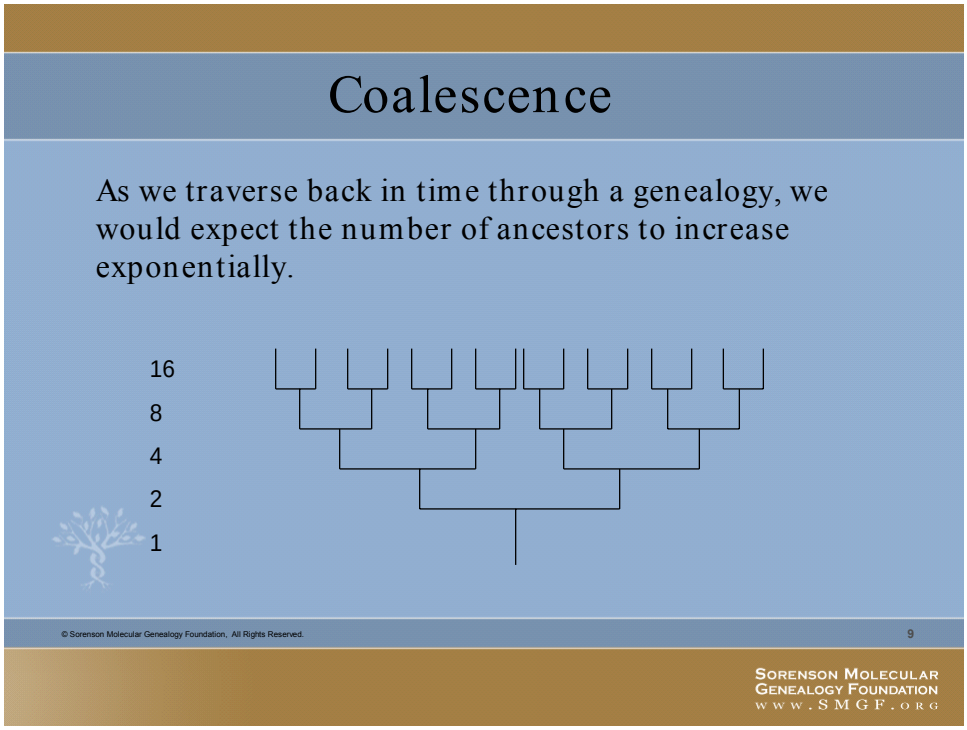
Living Individual

SORENSON MOLECULAR GENEALOGY FOUNDATION
WWW.SMGF.ORG

Of course, this would not be possible, as the world population estimate for the year 1250AD was approximately 450 million!! How do we explain this huge discrepancy? The answer is that we are all related and we share common ancestors much closer in time that we think. We all married our own cousins!!

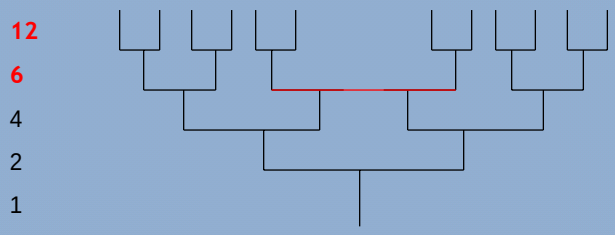


We are all related at different degrees. These family ties are continually lost at each generation. For example, few people know their third or fourth degree cousins, even though they share common ancestors relatively close in time.



Coalescence (continued)

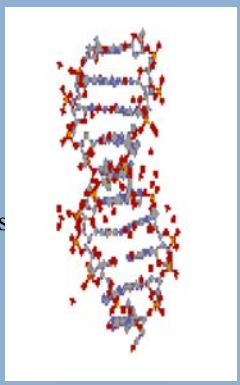
However, genealogies actually coalesce.



The reality is that the number of actual ancestors is much smaller than the number of possible ancestors. Our genealogy will expand and then coalesce.

3 major types of genetic data

- **Autosomal (Nuclear)**
  - Both males and females
  - Inherited equally from both parents
  - Recombination at each generation
  - >99% of your genetic information
- **Y Chromosome**
  - Males only
  - Follows paternal line <=> western surnames
  - 0.51% of total DNA
- **Mitochondrial DNA**
  - Both males and females
  - Maternal inheritance
  - 0.0006% of total DNA
- **[Also: X Chromosome]**



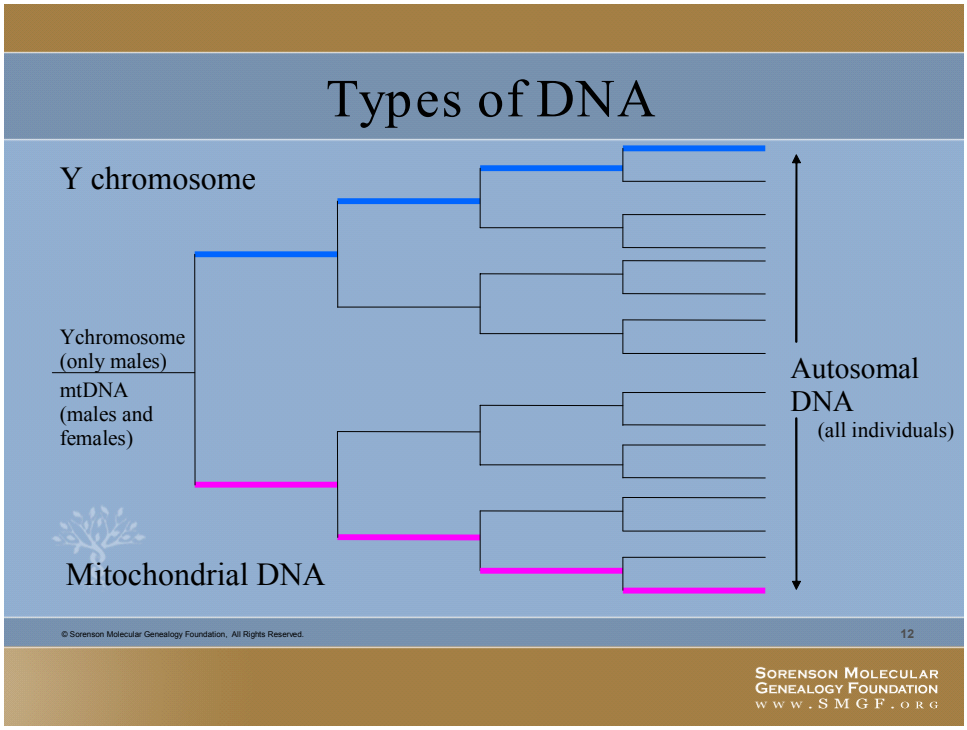
While many DNA inheritance patterns are complex and difficult to interpret, two methods of genetic analysis known as Y chromosome and mitochondrial DNA (mtDNA) testing are relatively straightforward. These types of analysis are currently

available to genealogists and are reliable for establishing certain types of family relationships. Other types of genetic analysis that could be used for genealogical purposes involve DNA from the autosomes (these are the remaining 22 pairs of chromosomes) and from the X chromosome (males inherit one from their mother, females inherit a copy from each parent).

The Y chromosome is very useful in establish paternal lineages, but it is limited in its use due the fact that only males carry it. It is inherited exclusively along the paternal line, from father to son, similar to the surname in many western (and even non-western) cultures. Two individuals sharing an identical, or very similar, Y chromosome profile would most likely share a common paternal ancestors within few generations. Based on how similar a Y chromosome is between two individual, it is possible to guess how many generations separate the two. Y chromosome has a higher rate of mutation than mtDNA. For this reason, Y chromosome testing is much more helpful to establish and verify relationships that occurred within the last 10-12 generations. This is a useful time frame in genealogical research because it is also the time in which surnames were in use. For example, two individuals sharing the same or similar surname may decide to have their Y chromosome tested to find out if they received their surname from a common male ancestor. This is particularly helpful when a surname is very common, thus leading genealogists on a search after the wrong ancestor. DNA testing on the Y chromosome could easily detect a specific lineage, even when there are two or more potential ancestors with the same name.

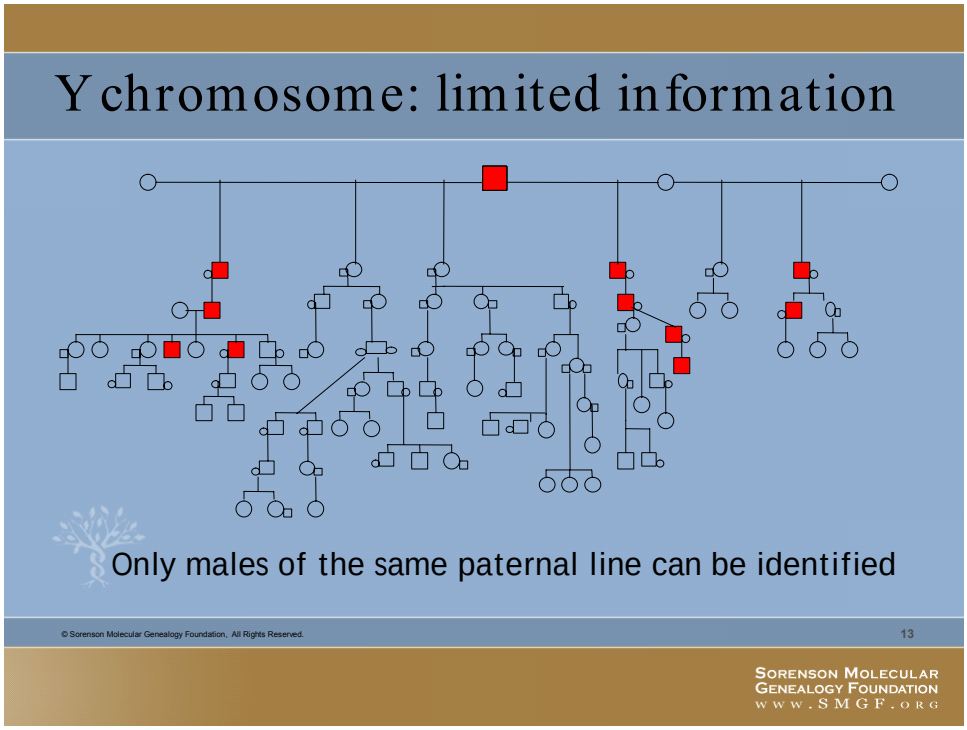
mtDNA is used to trace maternal lineages. It is also often used to trace ethnicity. A popular mtDNA testing in the USA is the Native American testing that allows people to verify if they had a maternal ancestor that was a Native American. mtDNA is more limited in genealogical applications because the surname along the maternal line changes at every generations and the mutation rate is much slower than the Y chromosome. This means that people sharing the same mtDNA profile might share a common maternal ancestors several hundred years in the past. Both Y chromosome and mtDNA testing are available commercially from a number of laboratories in USA and Europe. Also available are searchable databases that use these two genetic tests to discover biologically connections with others within the database. We will discuss one of these databases later on in this presentation.

Autosomal and X chromosome testing for genealogical purposes is still in its infancy. They are the focus of current academic and scientific research. Their inheritance patterns are much more complex to follow from generation to generation than the strict paternal Y chromosome and the strict maternal mtDNA, as shown in the next slide.



Although Y chromosome and mtDNA analysis can be extremely useful, they are limited to establishing strictly paternal or maternal relationships, represented by the outermost lines of a standard pedigree chart. A simple jump back 5 generations will

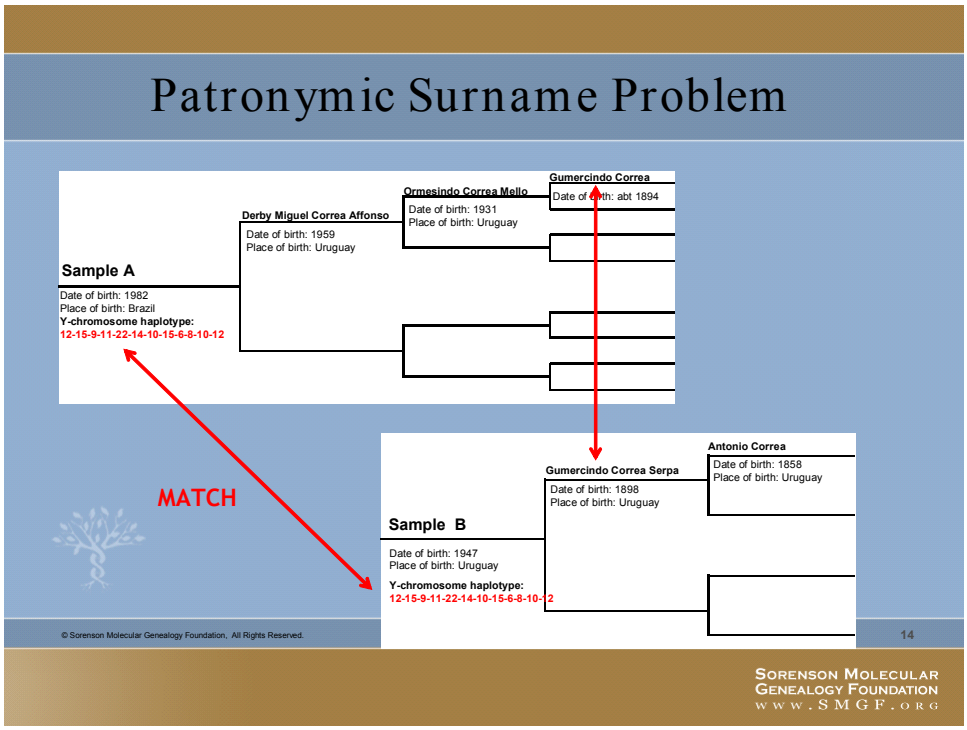
results in sixteen great-great-grandparents. However, Y chromosome and mtDNA testing will be helpful only in learning something about two of them. Over 99% of our DNA is inherited from all the other ancestors. This is the DNA found on the remaining 22 chromosomes (autosomes) and on the X chromosome. Testing these chromosomes will elucidate additional family relationships and provide a more complete picture of our genetic pedigree chart.



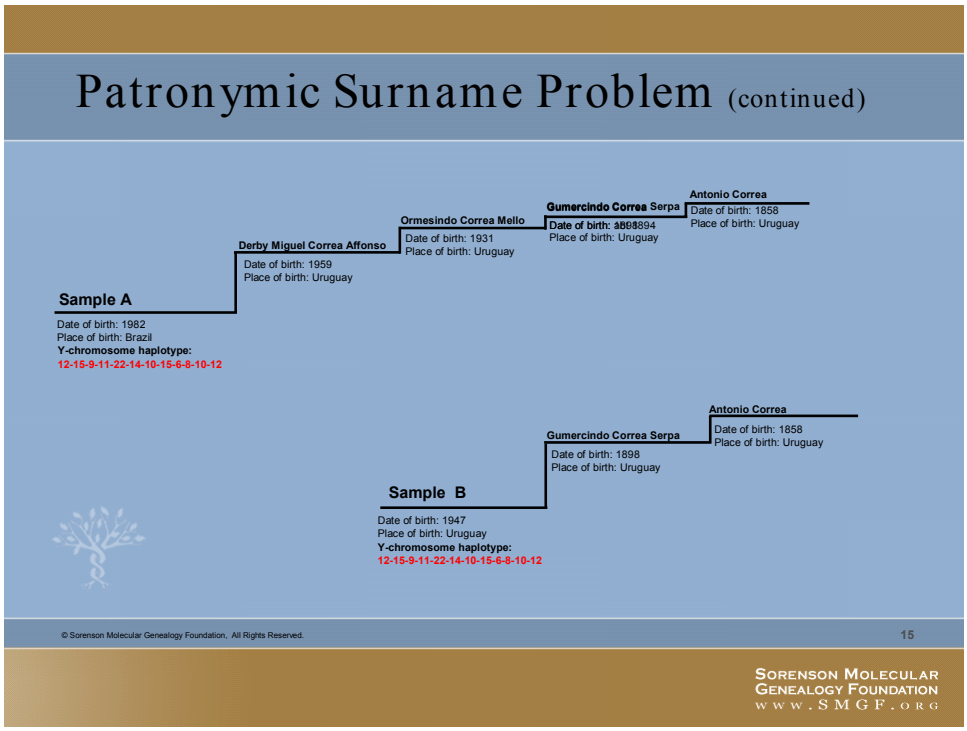
As explained earlier, the Y chromosome is only found in males. It is always passed from father to son. The Y chromosome remains essentially unchanged from one generation to the next. Because of these inheritance properties, Y chromosome testing can be a valuable tool for surname studies. Nearly two thousands are currently underway, with the number growing each week. Additionally, Y chromosome testing was used to support the existence of familial relationships in the highly publicized 1998 Jefferson-Hemings case and the Jewish priestly class of Cohen study.

References:

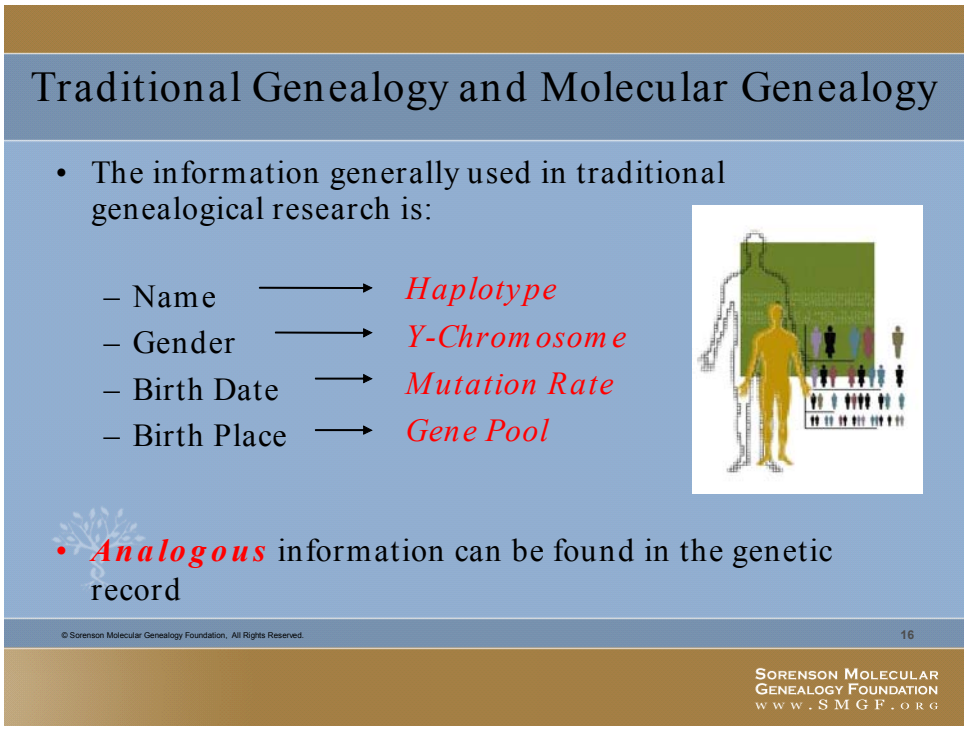
Mark A. Jobling. *In the Name of the Father: surnames and genetics*. TRENDS in Genetics, 17 No. 6 (2001): 353-357.
Chris Pomery. *Surname DNA Studies*, 2005 [online]. (http://www.dnaandfamilyhistory.com/documents/Surname_DNA_Studies.pdf). Accessed April 2005.
Eugene A. Foster, et al., *Jefferson fathered slave's last child.* Nature. 396 (1998): 27-28.
Karl Skorecki, et al.*Y Chromosomes of Jewish Priests*. Nature. 385 (1997): 32.



This is an example where Y chromosome testing was critical in linking two separate pedigree charts together. The information found in the two pedigrees was insufficient to confidently ascertain the uniqueness of a particular paternal ancestor. Y chromosome testing revealed that the two descendants (Sample A and Sample B) belonged to the same paternal line (meaning that they share a common paternal ancestor), even though the names and the birthdates were recorded differently.

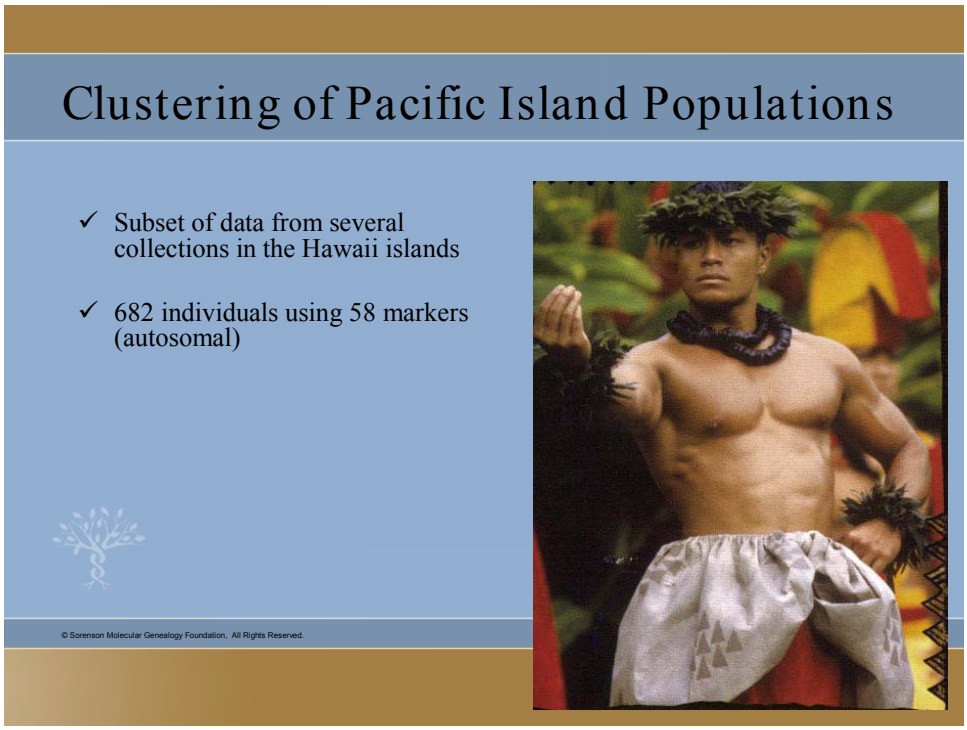


In addition, thanks to this genetic connection, Sample A can correct the information about his ancestor and add one extra generation to his pedigree chart.

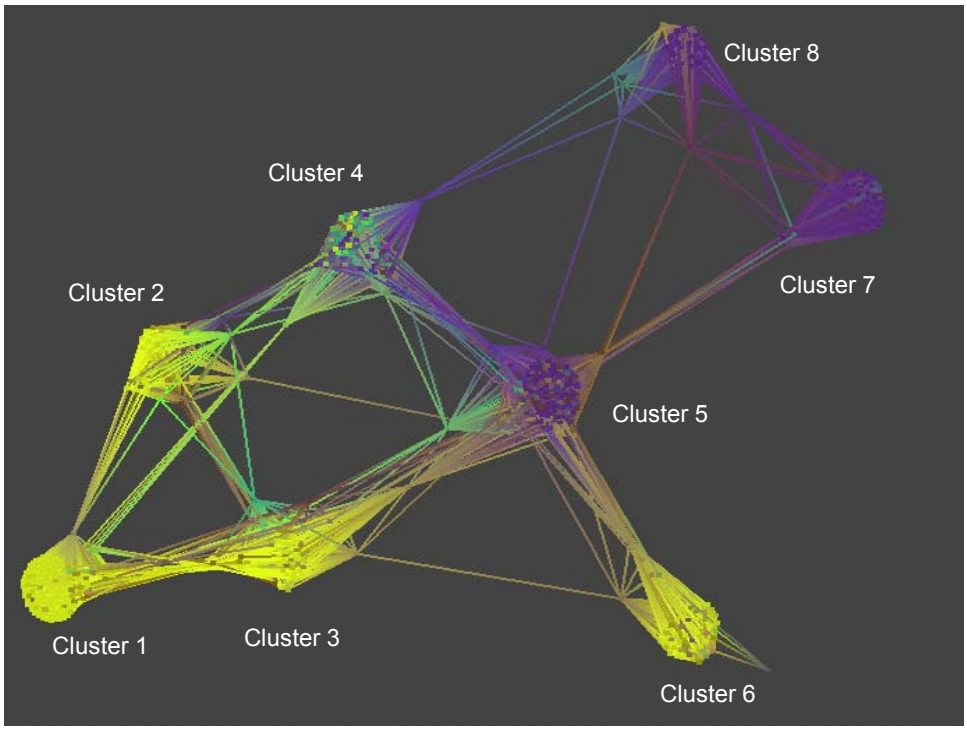


DNA contains the record of our family history. Specific information along the DNA strand are unique to a specific individual that lived in the past, that lives today, or that will live in the future. With traditional genealogy, researchers look for documents that would provide useful information that would lead to the identification of a unique ancestor. Molecular genealogy techniques allows the reconstruction of comparable information. The *haplotype* is a unique combination of genetic markers that would identify an individual or an ancestral lineage over others. The presence or absence of the Y chromosome would help in identifying the gender (males have the Y chromosome, females don't).

In addition, DNA change over time due to random mutations that occur between generations. These mutation can be taken into consideration to establish approximate rates that could be used to calculate and estimate the time frame when a particular ancestors lived. For example, two individuals sharing 36 out of 36 genetic markers on the Y chromosome would share a common ancestor within the past 2-6 generations; individuals sharing 35 out of 36 markers (due to a mutation on one of the markers) would share a common paternal ancestors up to 12 generations in the past; etc. Likewise, large correlated databases of genetic and genealogical information are helpful in assigning an individual (and his/her haplotype) to specific gene pools. The following preliminary study on populations from the Pacific area is helpful in better understanding clustering and assignment to geographic gene pools.



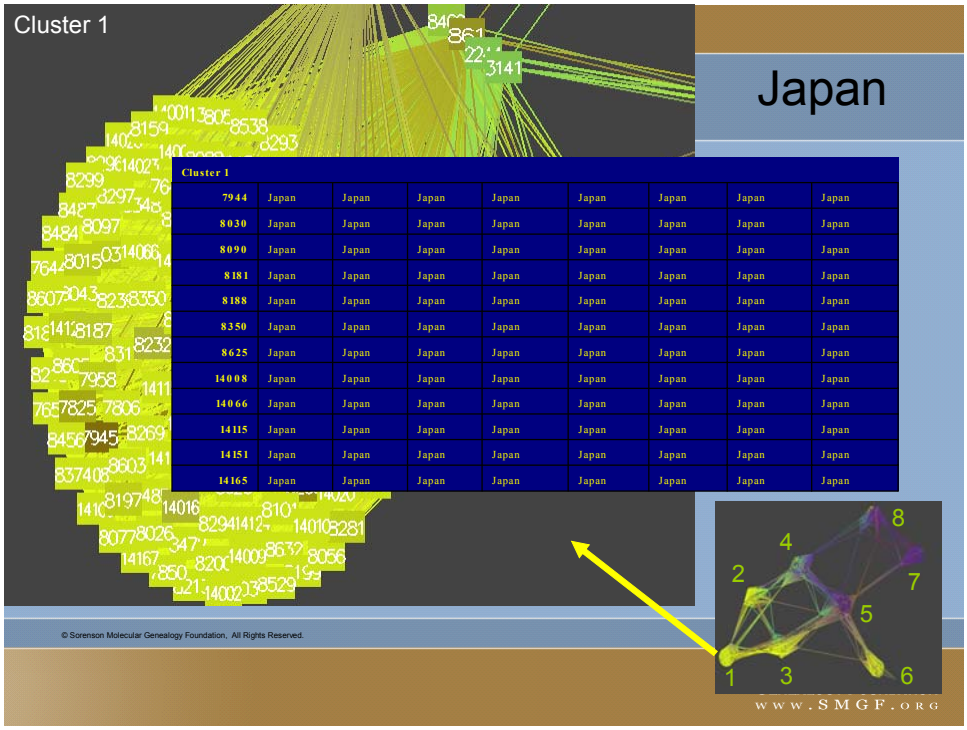
The following is an example of using a correlated database of genealogical and genetic data to create clusters (gene pools) where individuals sharing similar genetic markers could be assigned to.
682 individuals living in Hawaii submitted a DNA sample and a copy of their pedigree chart to the Sorenson Molecular Genealogy Foundation database. The DNA samples were analyzed at 58 autosomal markers. These are markers found in regions of the DNA that have been inherited from all of the ancestors and not just the paternal (Y chromosome) and maternal (mtDNA) lines. As many would know, Hawaii is a large melting pot of individuals from many parts of the Pacific. A sample population from Hawaii would represent many islands in the Pacific Ocean, as well as many other Asian and European countries.



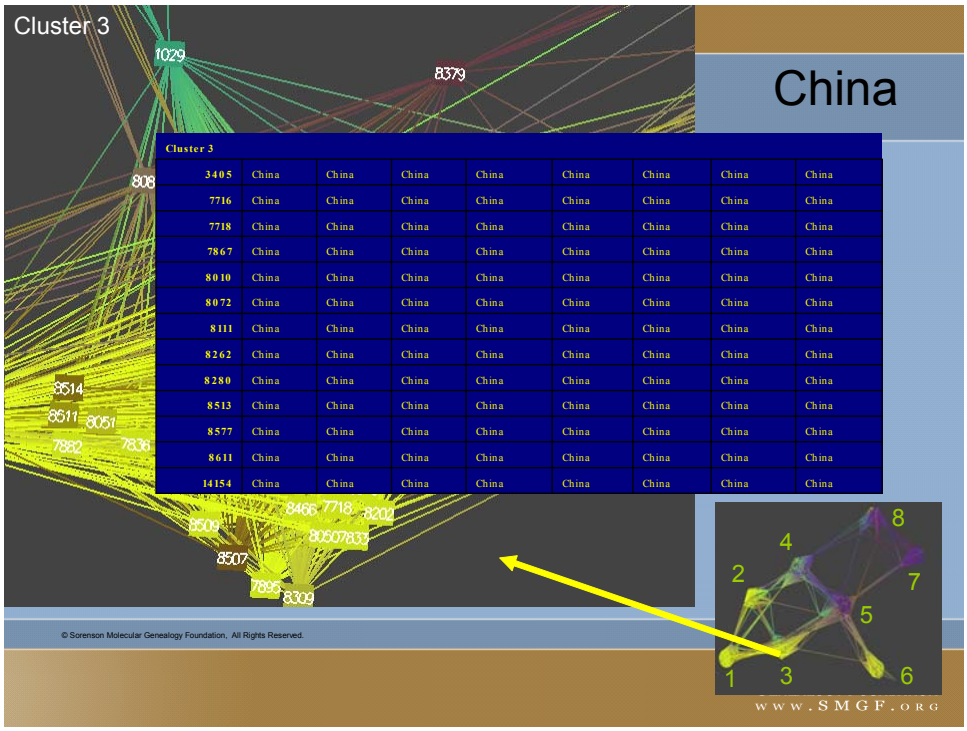
The samples were run using a program called STRUCTURE. This algorithm has been successful in identifying clusters at continental scales. This data set was clustered based solely on genetic data and compared with known genealogical data to group individuals into populations of origin within the relative near past (i.e. within the last 10 generations).
The DNA from the participants in these subset was analyzed at 58 markers on the autosomes (the first 22 chromosomes). STRUCTURE was run at k=2 through 20 with no prior genealogical or population data. This means that when we analyzed the DNA of these participants, we did not know anything about their genealogical information and therefore about their origin. After the DNA samples of the participants were assigned to a specific cluster, population assignments were made by extracting birthplace information from the genealogical data collected at the time of sampling. Genealogies varied in depth from 4 to 9 generations.
STRUCTURE output was used to generate TULIP (http://www.tulip-software.org) diagrams as shown in figure 1. Eight specific clusters representing the k=8 STRUCTURE run is depicted. Specific population assignments to the clusters were made by inspection of the genealogies of the individuals in each cluster. Percentages of parentage of each individual were then calculated for inclusion in the specific population assignments. For example, in the population labeled China, 93% of the birthplaces of terminal ancestors in the subset cluster were in mainland China. In the Samoan sample, 98% of the individuals terminal ancestral birthplaces were in Samoa. In the Hawaii cluster, only 79% of the terminal ancestral birthplaces were in Hawaii. Assignments therefore reflect two major components, the depth of the genealogical record, (the genealogical record ends prior to the 'actual' ancestral home of the individual) and the amount of admixture that is present in the Hawaiian sample (collection location).

References:

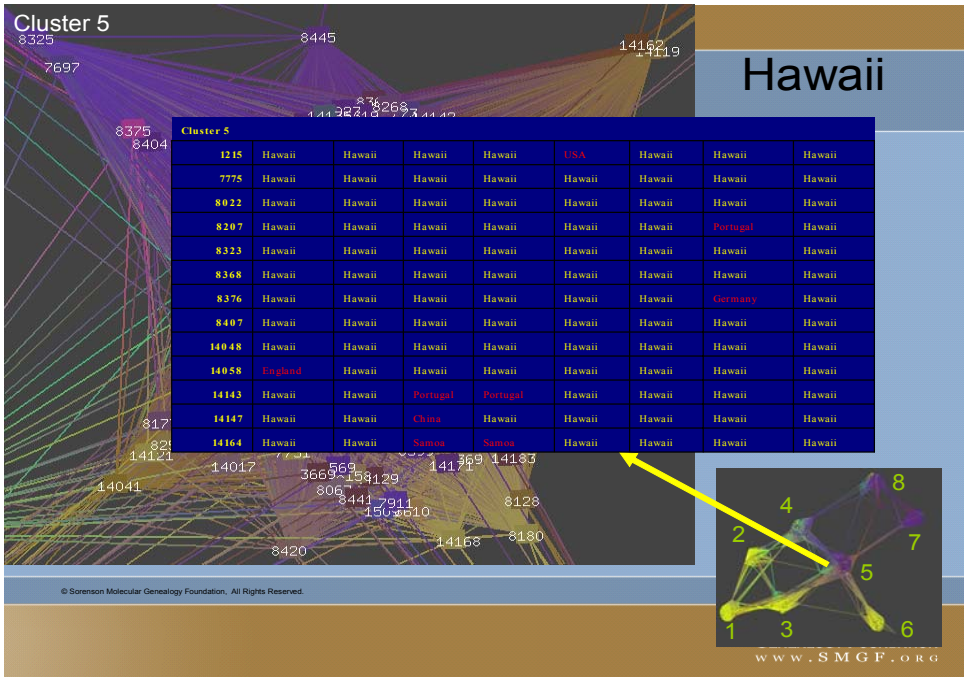
M. Bamshad, S. Wooding, W. S. Watkins, C. Ostler, M. Batzer, L. Jorde. 2003. “Human Population genetic structure and inference of group membership.” *American Journal Human Genetics* 72:578-589.
D. Falush, M. Stephens, J. K. Pritchard. 2003. ”Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.” *Genetics* 164:1567-87.
J. K. Pritchard, M. Stephens, P. Donnelly. 2000 “Inference of population structure using multilocus genotype data.” *Genetics* 155:945-59.



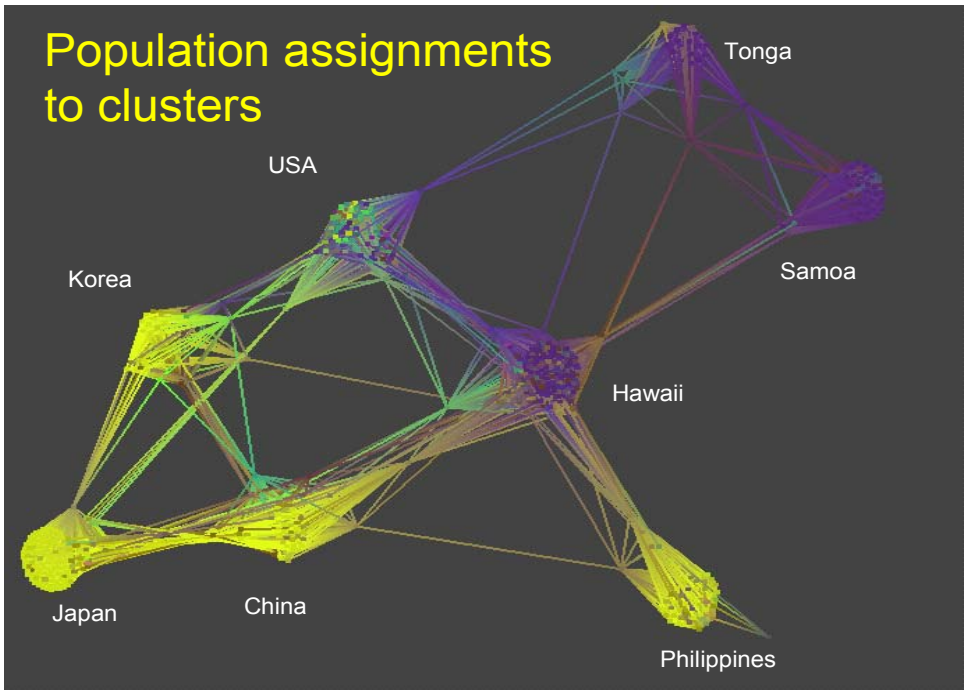
Based on the genealogical information submitted by the individuals in cluster 1, it was possible to assign this cluster to the country of Japan.



Likewise, cluster 3 was assigned to the country of China.



Cluster 5 was assigned to native Hawaiians, and so on.

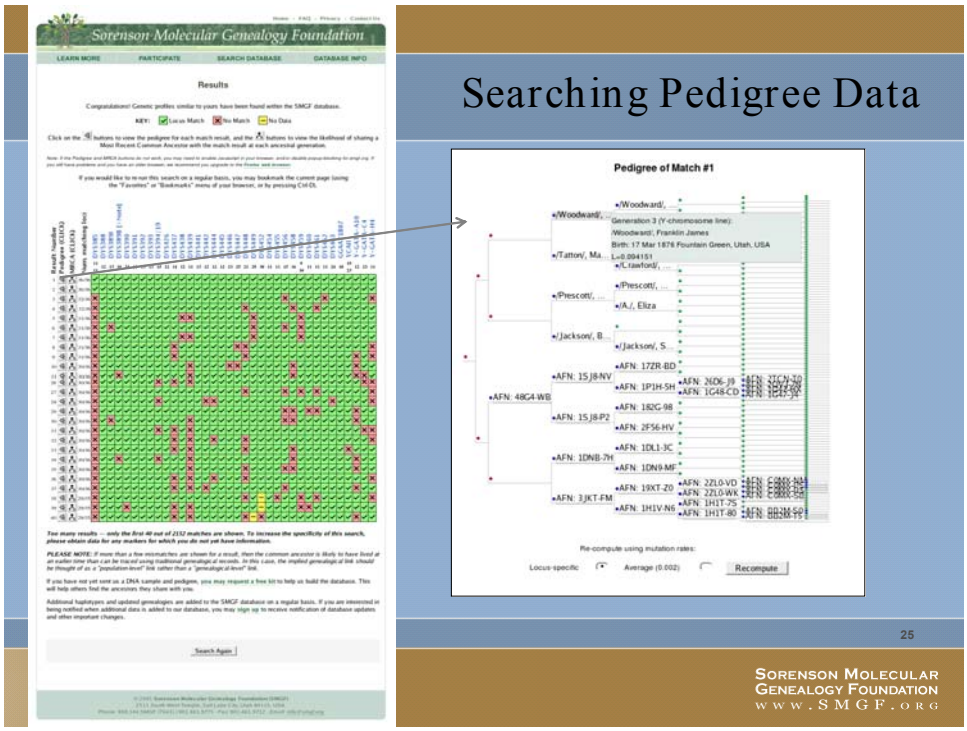


The cluster were created using exclusive genetic information with the algorithm STRUCTURE. Thanks to the available genealogical information, it was possible to assign each cluster to a specific geographic location. This preliminary study shows the potential of being able to assign an individual with unknown genealogical information to a specific ancestral cluster based on the information available in the correlated Sorenson Molecular Genealogy Foundation database. In other words, a large and comprehensive database of both genealogical and genetic data can be the answer to many in finding information about their ancestry based exclusively on the genes we all carry .

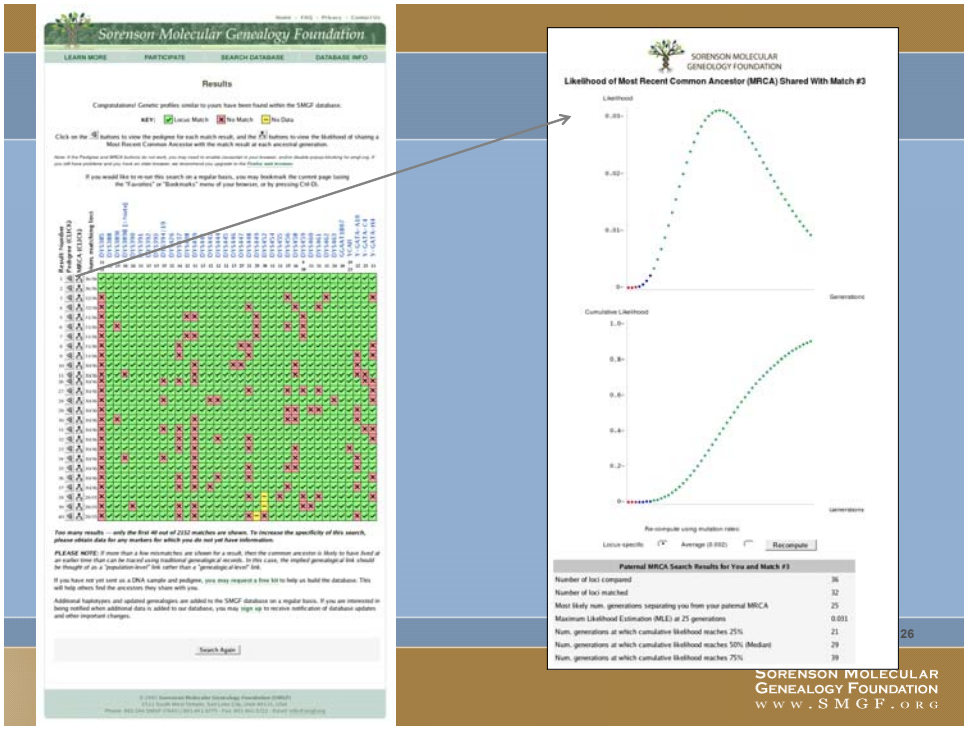
The Sorenson Molecular Genealogy Foundation has created an interactive website to fulfill three main objectives:

1. To be a source of information about molecular genealogy, where people can find educational material, scientific publications, examples, and other useful tools about the new and fascinating field of molecular genealogy;
2. To invite people to participate in the building of the world largest and most comprehensive database of correlated genealogical and genetic information. Those that would like to submit their DNA and genealogical information to SMGF can request a participation kit in the mail. Participation is free, confidential and voluntary;
3. To allow people to search our current Y chromosome database.

In March of 2004, the Sorenson Molecular Genealogy Foundation launched the first searchable database that uses the Y chromosome to retrieve genealogical information from people sharing the same, or similar, Y chromosome profile. This database is a subset of the larger database of over 50,000 DNA samples collected over the past five years by SMGF. In order to query SMGF Y chromosome database, individuals must first obtain their own Y chromosome profile by contacting an independent laboratory. Once the genetic data is obtained, it can be used to query the database, by selecting the corresponding values in the dropdown boxes on the query page. This database is updated with new data approximately every three month. The current version (April 2005) contains over 12,000 searchable Y chromosome haplotypes linked to ca. 500,000 ancestors. SMGF is currently working on additional searchable databases using both mtDNA and autosomal markers.



The database will display a list of exact and close matches to the Y chromosome haplotype submitted in the query page. By selecting the pedigree icon next to each match, a new window containing pedigree information will be displayed. Only genealogical data before 1900 is made available.



By selecting the MRCA calculation icon, statistical data based on the number of markers shared between the person making the query and the match from the database is displayed in a separate window.

The research conducted at the Sorenson Molecular Genealogy Foundation is possible thanks to the generosity of Mr. James L. Sorenson of Utah and to the thousands of people worldwide who have shared their genetic and genealogical information with us. We would like to extend an invitation to everyone to register their email address on our website and receive future updates about our work. Moreover, we would like to invite everyone to share their genetic and genealogical data with the purpose of building the world largest database of this kind and help reconstruct the family tree of humankind.