Introduction to Autosomal DNA

Blaine T. Bettinger, Ph.D., J.D.
www.DNA-Central.com
blainebettinger@gmail.com

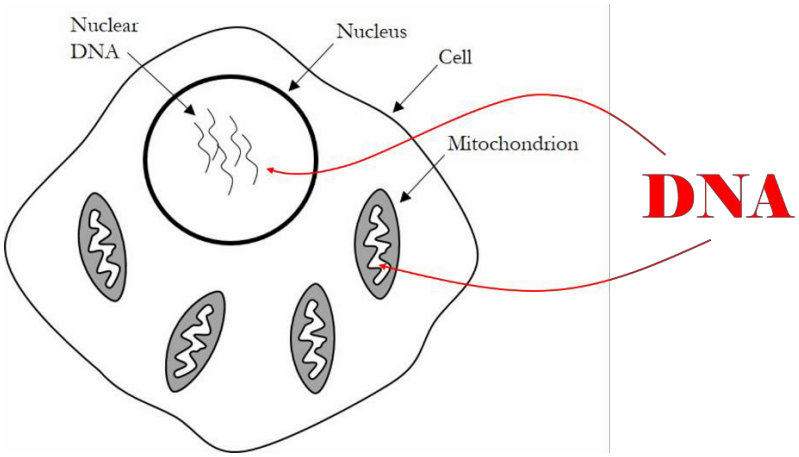
Introduction to DNA for Genealogists

Let's learn about some of the fundamentals of DNA testing. But don't worry, you won't need to remember anything from that high school or college biology class!

What is DNA?

DNA, or *deoxyribonucleic acid*, is a component of the cell (the basic unit of life) which carries the instructions for the development and operation of all living things. A piece of DNA is typically composed of two long intertwined molecules which are made up of smaller units called nucleotides. The two intertwined molecules interact to form a very long single structure called a *chromosome*. Chromosomes reside in the nucleus – or "control center" – of the cell. A typical human cell has *two copies* of each of 23 different chromosomes, for a total of 46 chromosomes. One set of each of the 23 chromosome pairs came from your father, and one set came from your mother.

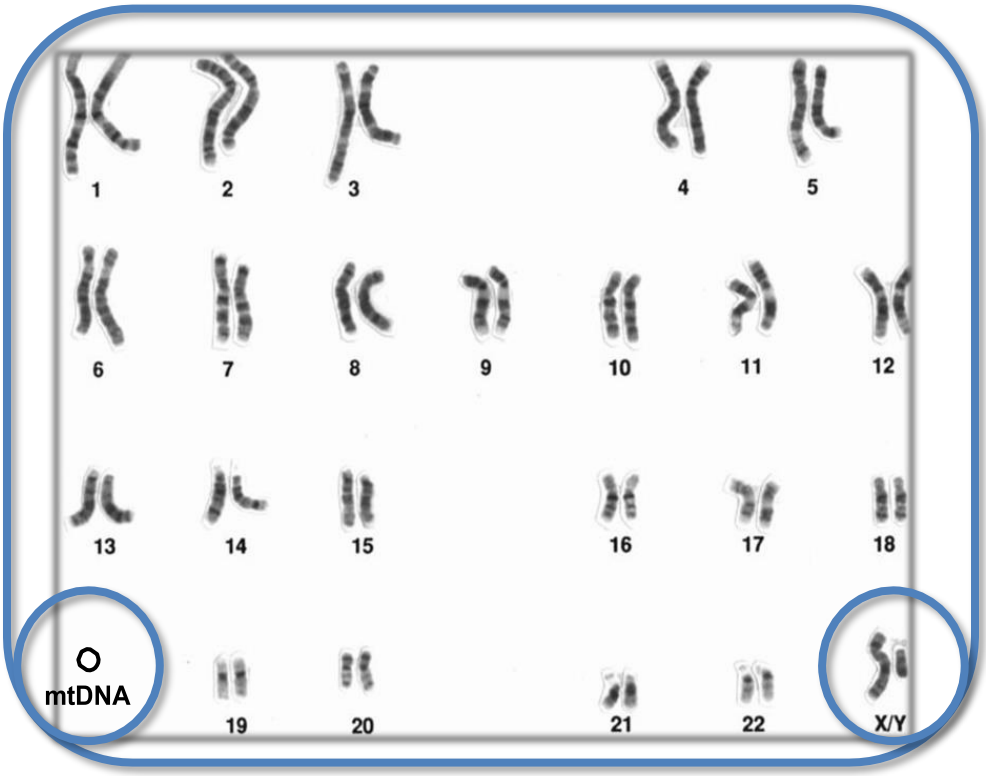
In addition to the DNA in the nucleus, there are hundreds or thousands of copies of a very small circular strand of *mitochondrial DNA* found in the many mitochondria outside the nucleus. These mitochondria are tiny powerhouses of the cell responsible for, among other things, creating the energy our cells need to function.



Types of DNA Tests

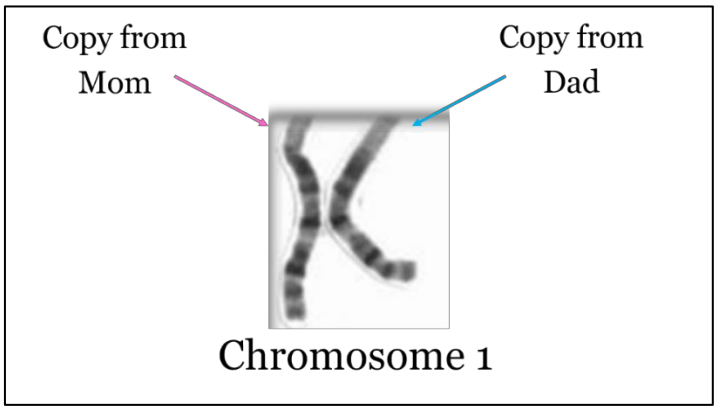
There are four different types of DNA tests for genetic genealogists:

1. Mitochondrial DNA (mtDNA) Test – This test analyzes the small circular piece of DNA found in the mitochondria, and will tell you about your direct maternal line (your mother's mother's mother's…mother).
2. Y Chromosome (Y-DNA) Test – This test analyzes the Y chromosome, a chromosome that is found only in males. As a result, this test will tell males about their direct paternal line (your father's father's father's…father).
3. Autosomal DNA (atDNA) Test – This test analyzes the 22 pair of non-sex chromosomes, including the copy of each chromosome you inherited from your mother and the copy you inherited from your father. As a result, these tests can tell you about both sides of your family.
4. X Chromosome (X-DNA) Test – This test analyzes the X chromosome, of which men have one copy (inherited from their mother) and women have two copies (one from their father, and one from their mother). Some autosomal DNA tests also examine the X chromosome.



Autosomal DNA Testing

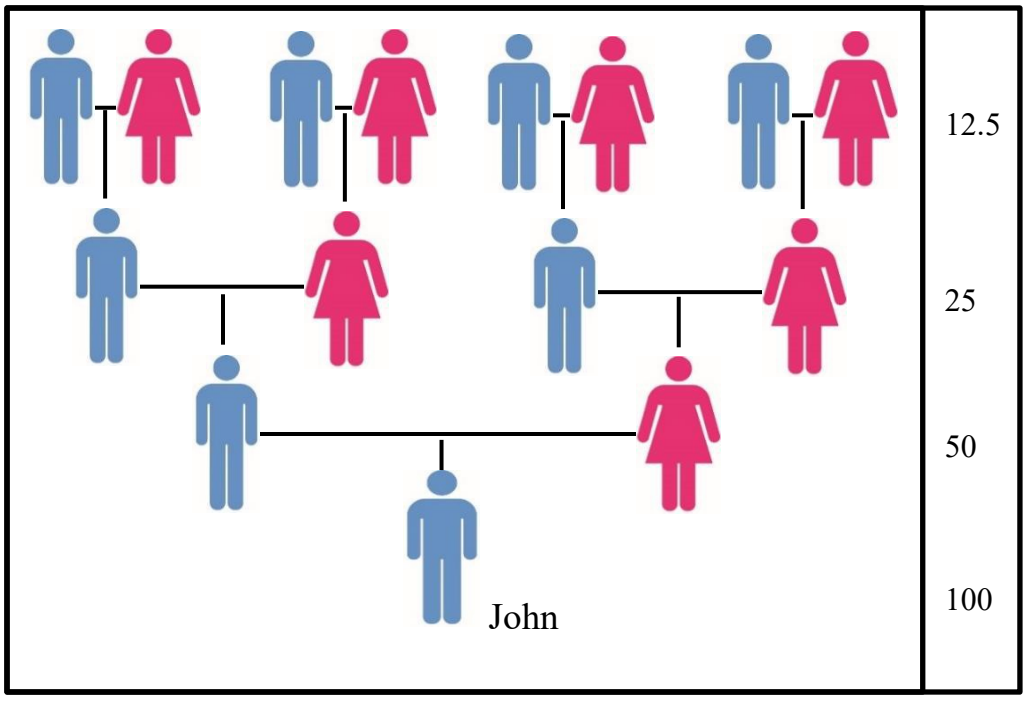
Autosomal DNA is the 22 pairs of non-sex chromosomes found within the nucleus of every cell. The 22 autosomes, or autosomal DNA chromosomes, are numbered approximately in relation to their sizes, with autosome 1 being the largest and autosome 22 being the smallest.



You have two copies of chromosomes 1-22, one copy that you inherited from your mother, and one copy that you inherited from your father:

Using Autosomal DNA

Autosomal DNA is the 22 pairs of non-sex chromosomes found within the nucleus of every cell. The 22 autosomes, or autosomal DNA chromosomes, are numbered approximately in relation to their sizes, with autosome 1 being the largest and autosome 22 being the smallest. The following figure follows the inheritance of autosomal DNA through four generations of a family, from eight great-grandparents to their great-grandchild John:

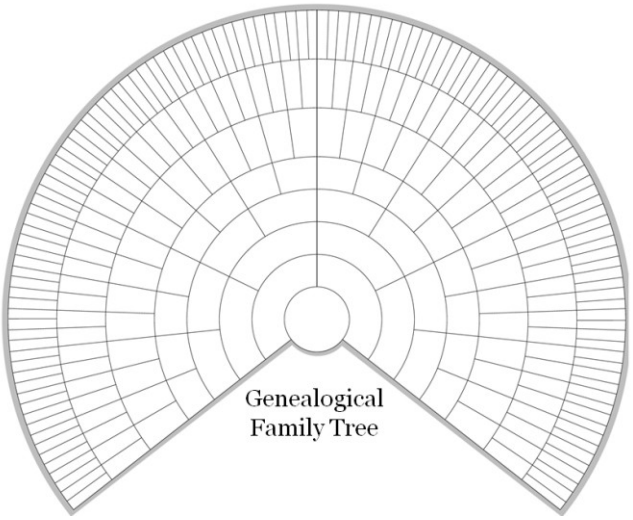


John received about 50% of his DNA from each of his parents, about 25% of his DNA from each of his grandparents, and about 12.5% of his DNA from each of his great-grandparents. Although not shown in this figure, Frank will inherit about 6.25% of his DNA from his great-great-grandparents, and so on.

1. You Have TWO Family Trees

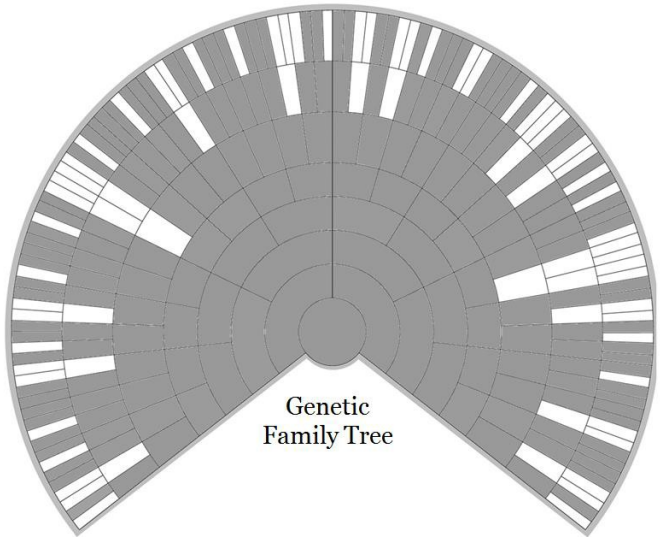
One of the most important aspects of genetic genealogy required to completely understand and interpret autosomal DNA test results is the fact that everyone has two very different, but overlapping, family trees.

The Genealogical Family Tree



The first family tree is your Genealogical Family Tree, which contains every ancestor that had a child who had a child who had a child, and so on, that ultimately led to you (see the figure below). This tree contains every parent, grandparent, and great-grandparent back through history. In most cases, this is the tree that genealogists spend their time researching, often using paper records such as birth and death certificates, census records, and newspapers to fill in. Many genealogists find that the paper trail ends or becomes much more difficult to identify beyond the 1800 or 1700's, making it difficult to fill in many of the openings in the Genealogical Family Tree.

The Genetic Family Tree



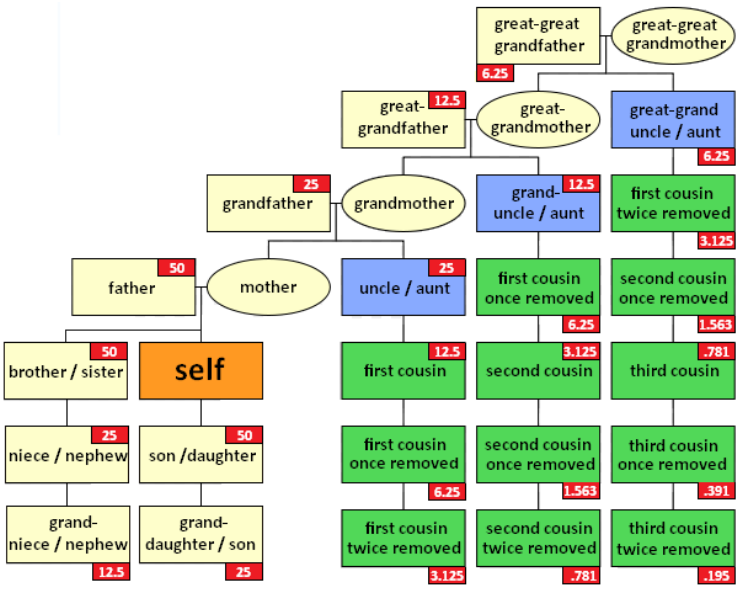
The second family tree is your Genetic Family Tree, which contains only those ancestors that contributed to our DNA. Not every person in your Genealogical Family Tree contributed a segment of their DNA sequence to your DNA sequence. A parent does not pass on all their DNA to their children (only about 50%); as a result, bits and pieces of DNA are lost in each generation. Somewhere between 5 and 7 generations back, your Genetic Family Tree starts to lose ancestors from your Genealogical Family Tree.

As shown in the figure below, your Genetic Tree is actually just a sub-set of your Genealogical Tree. Your genetic tree is guaranteed to contain both biological parents, who each contributed approximately 50% of your entire DNA sequence. Your genetic tree also likely contains each of your four biological grandparents and eight biological great- grandparents, but with each generation it is much less likely that every person in that generation contributed a piece of their DNA to your DNA.

2. Finding and Classifying Your Genetic Matches

Each of the testing companies return a list of genetic matches, which are all people in their database that share DNA with you above a certain threshold. The threshold used for matching is important; if the threshold is set too low you'll match everyone in the database. If the threshold is too high, you'll miss too many real matches.

Each company tries hard to find a suitable threshold, but it is important to keep in mind that all of the companies will provide matches that are "false positives" (matches who are not related to you in a genealogically relevant timeframe). This is just one reason that it is important to **concentrate on your best matches first**. Your "best matches" are those who share the most DNA with you.



The more segments you share with another person, and the larger those segments are, the closer your genealogical relationship with that person is. We inherit entire chromosomes from each parent, which are a collage of segments from our grandparents' chromosomes, which in turn are a collage of even smaller segments from our great-grandparents, and so on.

The chart below demonstrates the percentage of DNA that we share in common with our genealogical relatives. The more distant the relative, the fewer and smaller the segments of DNA that we share in common with that relative.

Although you are predicted to share 0.781% of your DNA with a third cousin, there's no guarantee that you will match a third cousin, as described in the next chart. Some (~10%) third cousins will not match. Anecdotally, however, no one to my knowledge has ever had a second cousin NOT match.

The amount of DNA shared by two people can help determine the genealogical relationship between those two people, although it is not a perfect predictor. For example, if you share 1500 cM with someone, that match is likely a grandparent/grandchild, aunt/uncle or niece/nephew, or half-sibling. However, if you share 75 cM with a match, it is not clear whether the match is a third cousin, second cousin once removed, or a more complicated relationship (e.g., multiple cousin).

For a more detailed chart, see ISOGG'S "Autosomal DNA Statistics" at (http://www.isogg.org/wiki/Autosomal_DNA_statistics):

Percentage	cMs Shared	Possible Relationship
100%	6766	Self, identical twin
50%	3400	Mother, father
50%	2640 - 3400	Full siblings
25%	1700.00	Grandfather, grandmother, aunt, uncle, half-sibling, double first cousin
12.5%	850.00	Great-grandparent, first cousin, great-uncle, great-aunt
6.25%	425.00	First cousin once removed
3.125%	212.50	Second cousin
1.563%	106.25	Second cousin once removed
0.781%	53.13	Third cousin

http://isogg.org/wiki/Autosomal_DNA_statistics

Chance of Sharing a Segment of DNA with a Genealogical Cousin			
	Family Tree DNA	AncestryDNA	23andMe
Closer than a Second Cousin	> 99%	100%	~100%
Second Cousin	> 99%	100%	>99%
Third Cousin	> 90%	98%	~90%
Fourth Cousin	> 50%	71%	~45%
Fifth Cousin	> 10%	32%	~15%
Sixth Cousin	< 5%	11%	<5%

The Shared cM Project

The Shared cM Project is a collaborative data collection and analysis project created to understand the ranges of shared centiMorgans associated with various known relationships. As of March 2020, total shared cM data for more than 55,000 known relationships has been provided. I am always collecting data, and perhaps the next update with have 100,000 relationships!

This March 2020 update is the third update to the original data, released in May 2015, and includes many thousands of new submissions.

Data Collection

Data was collected from participants using Google Forms, which collected the submissions into a spreadsheet. The Google Form contained data entry fields for required information ("Known Relationship," "Total Shared cM," "Number of Shared Segments," "Endogamy or Known Cousin Marriage" (YES/NO) and "Source" (23andMe, AncestryDNA, Family Tree DNA, MyHeritage, GEDmatch, or Other)), and optional data entry fields ("Longest Block," "Notes," and "Email Address").

A total of 59,714 submissions were made to the Shared cM Project as of 8 July 2019 (beginning March 4, 2015). For analysis, the submissions were downloaded as an Excel spreadsheet on 8 July 2019.

Initial Data Curation

Because "Known Relationship" was a text entry field, submissions varied considerably regarding the naming of various relationships. In this initial data curation stage, all decipherable relationships were converted to a uniform format (where "C" equals cousin and "R" equals removed). Submissions with indecipherable relationships were eliminated. Submissions with obvious data entry errors were also eliminated, such as those where the longest segment was longer than the total shared cM, or where there was text in the cM field instead of a number.

This initial data curation eliminated a total of 1,739 submissions (2.9%), bringing the total to 57,975 data points used for statistical analysis (although there were submissions included in this total for relationships not analyzed by the project).

A total of 48 different relationships ranging from Parent/Child to 8C were analyzed individually. The total number of submissions for each relationship varied, with a low of 33 for 5C3R, and a high of 5,281 for 2C1R.

Outlier Removal

Each relationship was analyzed individually, and obvious errors were removed (for example, 7 cM for a parent/child relationship). Then, a total of 1% of the submissions for each relationship was removed, removing 0.5% of the submissions at each end of the range. For example, if there were 200 submissions, 2 submissions were removed (the highest submission and the lowest submission).

Data Analysis

The dataset contained 55,418 submissions for the 48 different relationships analyzed by the project. Following outlier removal, the minimum, average, and maximum values of the remaining data points were identified for each relationship using standard methodology. Standard deviation was calculated using Excel.

For relationships where the minimum value was 0 cM shared, the averages were calculated only for cM amounts greater than 0 cM. *Accordingly*, these averages represent the average only for cousins sharing a detectable amount of DNA.

A histogram was created for each relationship. The histograms were created in Excel using the data for each relationship after outliers were removed.

Previous Versions of the Shared cM Project

	Launch Date	Total Submissions
Version 1.0	May 2015	>6,000
Version 2.0	June 2016	>10,000
Version 3.0	August 2017	>25,000
Version 4.0	March 2020	>59,000

Thank You

Thank you to EVERYONE that has submitted data to the Shared cM Project, whether one submission or many. YOU make this project possible!

For much more information:

- Bettinger, Blaine. "Version 4.0! March 2020 Update to the Shared cM Project!." (27 March 2020) https://thegeneticgenealogist.com/2020/03/27/version-4-0- march-2020-update-to-the-shared-cm-project/

The Shared cM Project – **Version 4.0** (March 2020)

Blaine T. Bettinger
www.TheGeneticGenealogist.com

How to read this chart:

Aunt/Uncle
1741
1201 - 2282

Relationship
Average
Range (min-max)

								Great-Great-Great-Grandparent		GGGG-Aunt/Uncle
							Great-Great-Grandparent		GGG-Aunt/Uncle	
Half GG-Aunt/Uncle 208 103 – 284	**Great-Grandparent** 887 485 – 1486							**Great-Great Aunt/Uncle** 420 186 – 713	**1C3R** 117 25 – 238	**2c3R** 51 0 – 154
Half 1C2R 125 16 – 269	**Half Great-Aunt/Uncle** 431 184 – 668	**Grandparent** 1754 984 – 2462					**Great Aunt/Uncle** 850 330 – 1467	**1C2R** 221 33 – 471	**2c2R** 71 0– 244	**3C2R** 36 0 – 166
Half 2c1R 66 0 – 190	**Half 1C1R** 224 62 – 469	**Half Aunt/Uncle** 871 492 – 1315	**Parent** 3485 2376 - 3720			**Aunt/Uncle** 1741 1201 - 2282	**1C1R** 433 102 – 980	**2c1R** 122 14 – 353	**3C1R** 48 0 – 192	**4C1R** 28 0 – 126
Half 3c 48 0 – 168	**Half 2c** 120 10 – 325	**Half 1C** 449 156 – 979	**Half-Sibling** 1759 1160 – 2436	**Sibling** 2613 1613 – 3488	**SELF**	**1C** 866 396 – 1397	**2c** 229 41 – 592	**3c** 73 0 – 234	**4c** 35 0 – 139	**5c** 25 0 – 117
Half 3c1R 37 0 – 139	**Half 2c1R** 66 0 – 190	**Half 1C1R** 224 62 – 469	**Half Niece/Nephew** 871 492 – 1315	**Niece/Nephew** 1740 1201 - 2282	**Child** 3487 2376 – 3720	**1C1R** 433 102 – 980	**2c1R** 122 14 – 353	**3C1R** 48 0 – 192	**4C1R** 28 0 – 126	**5C1R** 21 0 – 80
Half 3c2R 27 0 – 78	**Half 2c2R** 48 0 – 144	**Half 1C2R** 125 16 – 269	**Half Great Niece/Nephew** 431 184 – 668	**Great-Niece/Nephew** 850 330 – 1467	**Grandchild** 1754 984 – 2462	**1C2R** 221 33 – 471	**2c2R** 71 0– 244	**3C2R** 36 0 – 166	**4C2R** 22 0 – 93	**5C2R** 18 0 – 65
Half 3c3R	**Half 2c3R**	**Half 1C3R** 60 0 – 120	**Half GG Niece/Nephew** 208 103 – 284	**Great-Great-Niece/Nephew** 420 186 – 713	**Great-Grandchild** 887 485 – 1486	**1C3R** 117 25 – 238	**2c3R** 51 0 – 154	**3C3R** 27 0 – 98	**4C3R** 19 0 – 60	**5C3R** 13 0 – 30

Other Relationships
6C 18 0 – 71
6C1R 15 0 – 56
6C2R 13 0 – 45
7C 14 0 – 57
7C1R 12 0 – 50
8C 11 0 – 42

Minimum was automatically set to 0 cM for relationships more distant than Half 2C, and averages were determined only for submissions in which DNA was shared