Understanding Y-Chromosome DNA Testing

Bill Hough (wwhough@gmail.com) - 7/8/2014

There are just four chemical bases that make up the two coaxial spiraling strands, or double helix, of a DNA molecule, and they are usually known by the first letter of their chemical name. These abbreviated names are A, T, C and G. A always pairs with T on the opposite strand, and C always pairs with G. On each strand there is a very long sequence of bases, and because of the fixed pairing, when you know the sequence on one strand, you know the sequence on the other as well. Chromosomes contain DNA molecules, and the sequence of these four chemical bases on one of the two strands, called the positive strand, is what is studied in DNA tests. Specifically, the Y-chromosome is inherited basically unchanged by a son from his father, which is what makes the DNA in the Y-chromosome so important for determining paternal lines far into history. But there can be changes, or mutations. A mutation is a change in the sequence of the bases on each strand. It might be a change in a base chemical itself, or might be an insertion or deletion of a base or sequence of bases. These changes can occur when DNA is replicated to make new cells. Lots of new cells are produced when sperm is manufactured in a father, and that process is likely a major source of Y-chromosome DNA mutations from father to son..

There are two types of Y-chromosome mutations studied by the labs that do Y-chromosome DNA tests. There are roughly 500 regions on the Y-chromosome DNA strand that have been identified where short sequences of bases repeat. These are called “Short Tandem Repeats” or STRs. The DNA tests that have been around for a while looks at several of these regions on the DNA strand and counts how many times a sequence of bases is repeated. The “markers’’ in the FTDNA 37, 67, and 111 marker tests are names (usually beginning with DYS) for a few of the specific regions out of these roughly 500. The numbers in the FTDNA tables of surname or haplogroup projects under each "marker" are the count of the number of repeats of the base sequence at that marker. These counts, or number of repeats, can change or mutate relatively frequently. The possibility of a change varies with the particular marker (hence the colors of the column headings), the age of the father (probability of a mutation increases dramatically with the age of the father at conception), and probably other factors. By "relatively frequently," think generations. The more generations, the more mutations, or steps, in STRs you would expect. Maybe not expect, but consider a reasonable number that does not rule out the possibility that you have a recent common ancestor.

The other kind of mutation is an actual change of the chemical base at a position on the Y-chromosome. They have a long name you can look up, but it is abbreviated SNP (pronounced “snip”.) Until recently we thought of an SNP as being very rare, with times between them in a single line of descent being centuries, millennia, or more Your haplogroup identifies the sequence of SNP mutations from you to everybody's common ancestor (genetic Adam.) Haplogoups don't tell you a lot about recent common ancestors. But they can tell you something about the origin of your paternal line of descent, and if you have matching STR counts but are in different haplogroups, you can be sure you are not recently related.

Now since the probability of STR mutations in descendants of a common ancestor, particularly in markers with a low probability of mutation, is small even for a common ancestor that lived long ago, a member of a given haplogroup is likely to have many STR marker counts that match other members. That is what enables FTDNA to "predict" a haplogroup based on the STR marker counts. But if you look up that haplogroup, you may find there have been several other downstream (more recent) SNP mutations, or smaller haplogroups, and they haven't "predicted" those. The only way to see where you fall among the known downstream mutations is to have a Y-chromosome SNP test that looks for known mutations in your DNA. Safe to say, though, that if one of your known family group has an SNP test, the then "confirmed" haplogroup will almost certainly be the same for all of you. You can then join a haplogroup project, and find folks with different surnames that may be related. In my case, there are a lot of Scandinavian names, and it makes me think my ancestors came with the Viking invasion of England, which is where I know my more recent (15-1600) paternal line ancestors lived.

Two paragraphs ago, I qualified the statement that SNPs were considered very rare with "until recently." The recent change is that new SNP testing, for example FTDNA's Big Y or BritainsDNA Chromo 2, are now capable of looking at tens of thousands of positions in the Y-chromosome DNA molecule. Before the introduction of these new capabilities, the labs looked only at positions where SNPs were known to have occurred. The new tests are finding lots of previously unknown SNPs. Current tests are still nowhere near covering all the possibilities, as there are about 59 million base positions on one strand of Y-chromosome DNA. But we are probably close to the point that SNP mutations can be as useful as STR mutations in studies of recent common ancestry, the only problem being the cost of these new big SNP tests. For example, Max Huff and I are both 7 generations away from our known common ancestor who was born in Cheshire, England, about 1618, and with Big Y results we have identified 2 SNPs unique to my line and 3 unique to Max's. We also know that we have 1 step differences in repeat counts at 3 STR markers out of 37. Also these new tests have found branches of the haplogroup tree downstream of SNPs that have defined particular haplogroups for a while. These produce sub-haplogroups which are smaller (fewer people) than their parent. Think of a tree that keeps growing and producing new branches.

There are several editions of the haplogroup tree on-line (see, for example, the ISOGG Tree (http://www.isogg.org/tree/), and two ways to specify a specific haplogroup. The old way was with a string of letters and numbers, each letter or number indicating an SNP that divided the next upstream haplogroup. For example, Max and my haplogroup in the old way of designating them is currently I1a1b1, but in 2012 it was I1d1. This illustrates the problem with this naming system. They both meant that we had a chain of mutations from genetic Adam that had been identified, and if you look at the 2014 and 2012 trees on the ISOGG web site, you will see that they both ended with SNP P109. What happened is that two additional branch SNPs were identified in 2013 that were between Adam and P109. So the new way of naming a haplogroup is first by the major division of Adam, in our case I, followed by the last SNP that differentiates us from everybody else, in our case (until Big Y), P109. The syntax of the new naming system is then I-P109+, where the "+" indicates positive for the SNP P109. With Big Y results, we recently learned that we have 3 previously named SNPs below P109, and belong to a more definitive haplogroup that we can call I-S14887+. We also know that there are 12 newly identified SNPs below S14887 that Max and I share, but we don't know the order in which they occurred between our earliest ancestor with S14887 and our most recent common ancestor, that guy in Cheshire 7 generations ago. Then there are 2 unique ones that I show and 3 unique ones that Max shows, which are called Private SNPs. Private SNPs appear in only one family.

Now what do we mean when we say that we are positive for SNP P109. P109 is the name given to the fact that at position 15426005 on the DNA molecule in our Y-chromosome the chemical base T, which genetic Adam and every other male except those with the P109 mutation have at that position, mutated to a C in someone who lived about 3000 years ago, and all his paternal line descendants, like Max and I, have a C instead of a T at that position. On the same ISOGG web site, you can find in the "Index to Y-DNA SNPs" the position number and change in the base chemical for most named SNPs discovered before 2013. The base chemical everybody, including Adam, had before the SNP occurred is called the "reference," and the mutated chemical base is called the "variant." These are also sometimes called "ancestral" and "derived" respectively. By definition, genetic Adam had no variants. FTDNA uses the term "novel variants" in their Big Y reporting, and gives a position, the reference (usual chemical base, and the variant (mutated chemical base) for that position. These include those SNPs or variants which do not yet have SNP names including Private SNPs which may never be named.

So far we have talked only about Y-chromosome mutations, which are important for tracing the strictly paternal line. There are other tests, 23andMe, FTDNA family finder, and I think probably the Geo 2.0 test, that look at the non-sex (not the X and Y) chromosomes or other DNA. They can give you clues, maybe even knowledge, about family connections that are not dependent on strict paternal or maternal lines. They are based on the fact that in your non-sex chromosomes, you inherit half your DNA from you mother and half from your father.