What Is This DNA Stuff and Why Should I Care?
Part 2: DNA Testing Procedures

David R. Neithamer

Now that you have a little background on the structure of cells, where DNA is located within the cells, and the structure of DNA itself, let's discuss how the 4 major testing companies (Ancestry.com, My Heritage, 23andMe, and FTDNA) actually test your DNA sample. What happens, after spitting in the tube or swabbing your cheek, during those 4 to 6 weeks while you wait to receive your results? For most of us it is a mystery, but how they do it is almost as interesting as the final results they provide.

It turns out that all of the testing companies now use the same basic technology to analyze your DNA - microarray beadchips, all produced by the same company (Illumina). MyHeritage, Ancestry.com, and FTDNA, use the Illumina OmniExpress beadchip whereas 23andMe uses the new Illumina Infinium® GSA beadchip. Although all of the beadchips come from the same company, the tests are not created equal. The beadchip is only one part of the process. Each company develops their own algorithms, analysis and matching criteria, population references, and reports. Furthermore, each company offers different tools for you to analyze your data, creating variations in results, accessibility, and usefulness. Before we discuss microarray beadchips any further, let me describe how they prepare your sample to be tested.

After you spit into the tube or swab your check you immediately seal it with a stabilizer solution. This multicomponent solution stabilizes the pH, protects the sample from thermal temperature swings that it may experience during shipping, protects the sample from degradation, and prevents bacterial growth in your sample. You don't want your DNA mixed up with bacterial DNA. Once your sample is received, they need to isolate the DNA from the cheek cells in your sample. The membranes of cells are made up of lipids, also known as fats. Soap is very good at breaking up and dissolving fats. Therefore, when you mix soap with the cells, it basically breaks up the membrane of the cell, thus causing the DNA to separate out. The soap does not interfere with the DNA because DNA does not contain lipids. The sample is then centrifuged to separate the insoluble cell parts from the DNA containing solution.

Now that they have an isolated sample of your DNA, the testing procedure begins. The assay only requires about 0.2 mg of the original DNA sample. That's about half of the weight of an average grain of sand and contains about 1 million to 10 million DNA molecules! Although this may seem like a lot, it is not enough to proceed with testing, so they need to amplify the DNA by making numerous copies of the DNA strands.

Figure 1 shows the six steps used to prepare the sample. First **(1)**, the DNA sample is thermally denatured and neutralized at about 95°C for about 1 minute to separate the DNA double helices into single strands. Reagents are then added to the sample including the four nucleotide bases (A, T, G, and C) to prepare them for amplification. Second **(2)**, using a procedure called the polymerase chain reaction (PCR) they make numerous copies of the DNA by cycling the temperature of the sample between about 70°C and 95°C with each cycle requiring about 5 minutes. After about 25 to 35 cycles, the sample is cooled to about 70°C to be annealed for about 5 to 10 minutes, and then cooled to 4°C until the DNA is ready for the next step. The PCR amplification reaction is highly automated and is usually run overnight on a Thermal Cycler PCR machine. The PCR reaction typically increases the number of DNA molecules by a factor of about 1 million to10 million, depending on the number of cycles used.

Part 2: DNA Testing Procedures

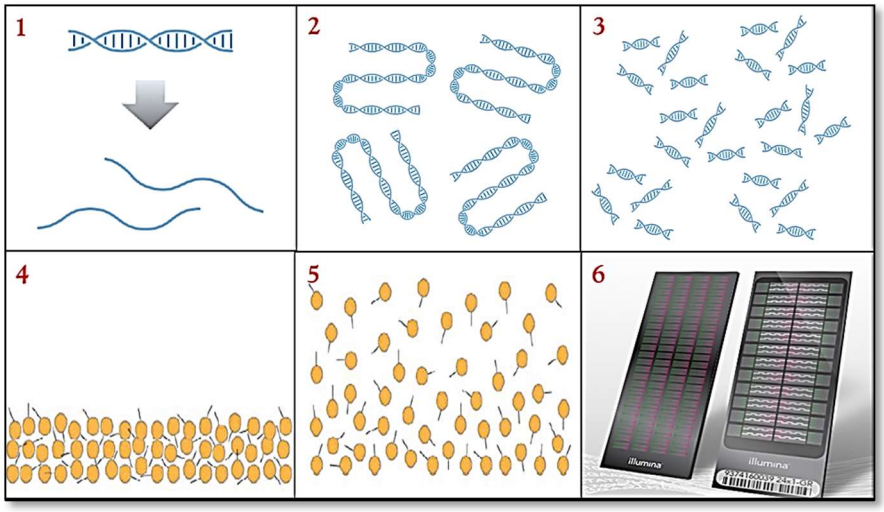

Figure 1. DNA Sample Preparation
Source: www.illumina.com

The DNA sample then needs to be chopped into smaller fragments which can then be analyzed **(3)**. This is accomplished using a controlled enzymatic process which cuts the DNA into fragments like cutting a string multiple times with scissors. In step four, the fragmented DNA strands are then precipitated with isopropanol (rubbing alcohol), centrifuged, and washed at 4°C several times **(4)**. The purified DNA fragments are then re-suspended and denatured in an ionic salt buffer solution which chemically interrupts the hydrogen bonding interactions between the nucleotide base pairs along the DNA fragment strands **(5)**. Denaturing effectively splits the double helical fragments into 2 single DNA fragment strands. Finally, the purified, denatured mixture is added to a microarray beadchip and incubated overnight in an oven to hybridize or attach the DNA fragments to the microarray beadchip substrate **(6)**.

The microarray beadchip itself is an amazing piece of technology. The beadchip shown below in **Figure 2** contains 12 sections which can be used to test 12 different individuals **(1)**. If you look at the beadchip under a microscope **(2)**, you will see that there are approximately 750,000 etched wells, with each well containing a tiny silica microbead **(3)**. Each microbead is about 1.5 microns (1/17,000 of an inch) in diameter.

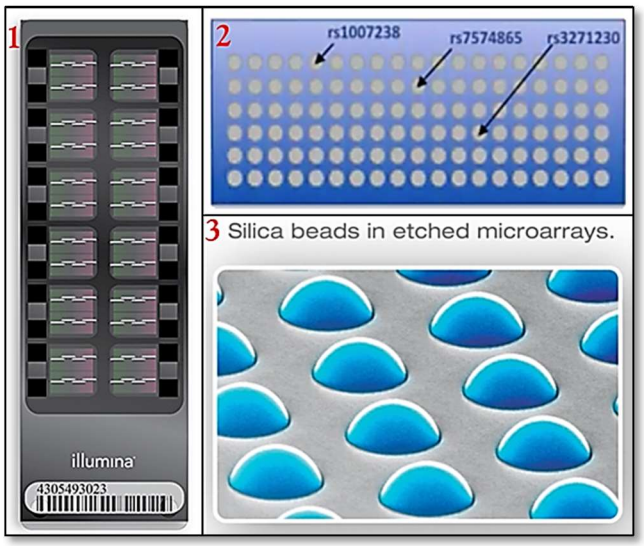

Figure 2. Microarray Beadchip
Source: www.illumina.com

Part 2: DNA Testing Procedures

Each individual microbead has been modified with many attached copies of a synthesized nucleotide sequence about 50 base pairs long. Each microbead contains a different nucleotide sequence that will be used to hybridize or attach the DNA fragments and to test for a specific nucleotide (A, T, G, or C) at a specific location along the human DNA strand. **Figure 3** shows this hybridization process. In panel **(1)** we see the modified microbead containing the short nucleotide sequences dangling from the surface. When the fragmented DNA solution is added **(2)** and annealed in an oven, the complementary fragments of DNA that exactly pair with the synthesized nucleotide sequence attach to the microbead **(3).** If we zoom into the surface of the microbead **(4),** we see the DNA fragments attached to the microbead sequences on the surface. In this case, we are interested in the first nucleotide from the DNA sample fragment that is not bound to the microbead-containing sequence. I'll call this the first dangling nucleotide and in Figure 3 it is a G. This is also called a single nucleotide polymorph (SNP).

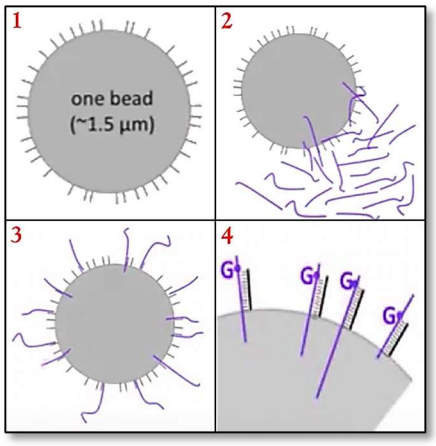

Figure 3. Hybridization of DNA onto the Beads
Source: www.illumina.com

Two double helical strands of DNA are derived from the chromosomes. One double helix strand originates from the father and one complimentary double helix strand originates from the mother. Each double helix has a single forward (+) strand and a single reverse (-) strand. **Figure 4** shows the ladder view of two 15-sequence fragments **(1)**, one from the father's DNA (blue) and one from the mother's DNA (red) and the forward (+) and reverse (-) strands are shown for each. When the DNA fragments are denatured prior to hybridization, this would lead to 4 single strands of DNA that could potentially be tested. However, when the sample is hybridized onto the microbeads, only the fragments derived from the forward (+) strands attach to the beads because the surface fragment nucleotide sequences on the microbeads were designed specifically to capture only the forward (+) strand fragments. The fragments from the reverse (-) strands are completely washed away. The right side of Figure 4 shows the complimentary forward (+) strands from the father and mother, for our 15-squence example **(2)**. Complimentary nucleotides (nucleotides that are found at the same specific locations along the DNA strands from the father and mother) are called alleles. If the alleles have the same nucleotide bases, they are called homogeneous alleles (examples of three are circled in yellow); if the alleles have different nucleotide bases, they are called heterogeneous alleles (examples of three are circled in maroon). The testing companies report the allele values at specific SNP locations along the DNA strand and report these in the raw data from your DNA test, although they cannot identify whether they came from the father's or mother's DNA.

Part 2: DNA Testing Procedures

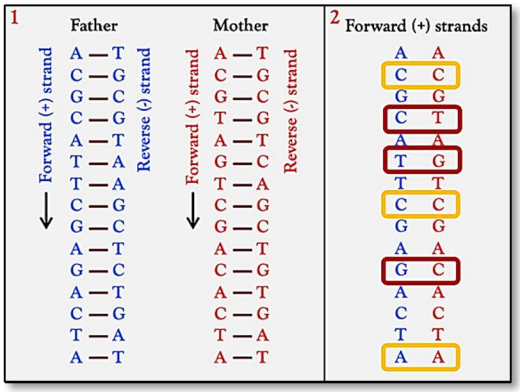

Figure 4. Denaturation and Separation of Forward (+) Strands
Source: Generated by the author

We now understand how the DNA is fragmented, purified, and hybridized onto the individual microbeads on the microarray beadchip, but how do they actually determine the value of the first dangling nucleotide (A, T, G, or C) from each microbead? In short, they use a modified, fluorescence-labeled nucleotide tag which binds to the last dangling nucleotide but not any other nucleotides along the dangling chain. A modified fluorescence-labeled nucleotide tag is a nucleotide base (A, T, G, or C) that has been modified with a blocking group that prevents it from reacting anywhere along the DNA fragment strand except with the corresponding last dangling nucleotide. Additionally, and it also contains a molecule that shines a particular color light when irradiated with a laser. This modified molecule is called a tag and we label the tags A*, T*, G*, or C*.

Figure 5 shows three examples of how the allele identification works. In **1A** a hybridized microbead has only an A as its last dangling nucleotide. This arises from a homogeneous allele in which both the mother and father have an A at this particular location in their DNA strand. In **1B** the microbead is annealed with a mixture of tagged molecules (A*, T*, G*, or C*), but since A only pairs with T, only the T* binds to this microbead. Finally, in **1C**, T* emits red light when irradiated with a laser and a computer counts this location as being identified as a homogenous allele pair A-A because only A was present. Similarly, in **2A** a hybridized microbead has only a G as its last dangling nucleotide. This arises from a homogeneous allele in which both the mother and father have a G at this particular location in their DNA strand. In **2B** the microbead is annealed with a mixture of tagged molecules (A*, T*, G*, or C*), but since G only pairs with C, only the C* binds to this microbead. In **2C** of this example, C* emits green light when irradiated with a laser and a computer counts this location as being identified as a homogenous allele pair G-G since only G was present.

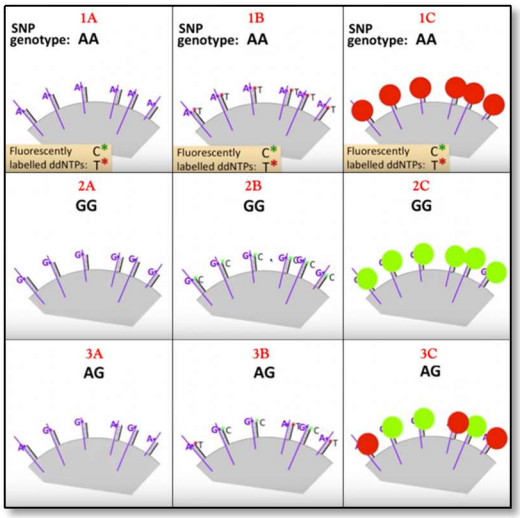

Figure 5. Fluorescence Labeling to Determine Allele Values
Source: www.illumina.com

Finally, in **3A** a hybridized microbead has both an A and a G as its last dangling nucleotides. This arises from a heterogeneous allele in which the mother has an A and the father has a G at this particular location in their DNA strand or vice versa. In **3B** the microbead is annealed with a mixture of tagged molecules (A*, T*, G*, or C*), but since both A and G are present, both T* and C* bind to the microbead. In **3C** of this example, when both T* and C* are present, the microbead emits a combination of red and green light, that is perceived as yellow when irradiated with a laser, and a computer counts this location as being identified as a heterogenous allele pair A-G. Again, it cannot identify which nucleotide came from the mother or father, only that one came from one person and the other came from the other person. As mentioned earlier, there are about 750,000 microbeads on the Microarray beadchip and each one is analyzed with a laser and individually counted by computer. Each microbead then gives a value of one single nucleotide polymorph (SNP) in your DNA testing results.

Figure 6 shows the results of the first 20 SNPs from my Ancestry.com test. The rsid (Reference SNP cluster ID), chromosome, and position values correspond to the genetic identification number of a particular SNP, the chromosome that the DNA came from, and the specific location that the SNP is located along the DNA strand of that chromosome. The two allele values are the two SNP values at that location. The homogeneous alleles are shown in light blue and the three heterogeneous alleles are highlighted in yellow. From this test alone I cannot determine which allele came from which parent. However, if my mother or father tested their DNA, then by comparing my SNPs with theirs at the same location, I could determine which allele came from which parent.

rsid	chromosome	position	allele1	allele2
rs369202065	1	569388	G	G
rs199476136	1	569400	T	T
rs190214723	1	693625	T	T
rs3131972	1	752721	G	G
rs12562034	1	768448	G	G
rs115093905	1	787173	G	G
rs6681049	1	800007	C	C
rs28444699	1	830181	A	A
rs4970383	1	838555	C	C
rs4970382	1	840753	T	T
rs11516185	1	843405	A	G
rs4475691	1	846808	C	C
rs13303369	1	852875	T	C
rs4970461	1	852964	T	G
rs7537756	1	854250	A	A
rs7418179	1	858801	G	G
rs13302982	1	861808	G	G
rs1110052	1	873558	T	T
rs7523549	1	879317	C	C
rs2272756	1	882033	G	G

Figure 6. The First 20 SNP Values from Chromosome 1 of My Ancestry.com DNA Test
Source: Raw DNA data file for the author downloaded into MS Excel from Ancestry.com

You might wonder how this test procedure is different from The Human Genome Project (HGP). The HGP was an international scientific collaborative research project with two main goals: 1) to determine the sequence of the 3 billion nucleotide base pairs that make up human DNA and 2) to identify and map all of the genes of the human genome. The HGP launched in 1990 and was "completed" in 2003, but genome "finishing" work continued for over a decade. The sequence is a mosaic of 5 anonymous individuals selected from a number of donor samples. Sequencing was performed in 20 research centers in the U.S., U.K., Japan, France, Germany, Spain, and China and the publicly funded project cost around $3 billion to complete. The main benefits to arise from this endeavor are to better understand the underlying causes of diseases, to help develop better treatments and medicines, to further anthropologic studies and to improve forensics testing. The HGP actually identified all 3 billion nucleotides that make up our genome. Knowing the complete genome allowed for the development of the microarray beadchip technology. The DNA testing companies only test approximately 750,000 SNPs out of a total 3 billion base pairs, or 0.025% of our DNA. This may seem like very little, but it is enough to give us great insights into our past and help us to identify genetic cousins.

Since we mentioned genetic cousins and genealogy, it is important to know that you have two different family trees - your genealogical family tree and your genetic family tree. As shown in **Figure 7** on the left, a 7-generation genealogical tree contains 254 ancestor names. If you add 3 more generations, a 10-generation

genealogical tree would contain 2046 names. This is the tree that you spend most of your time researching and populating with family names. The relevant timeframe for such a tree is the present-500 years ago because, in most cases, the paper trail runs out 500 years back or sooner.

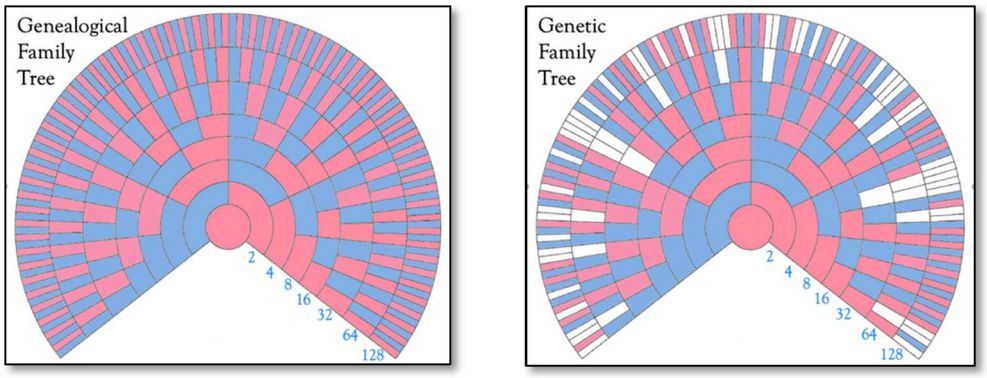

Figure 7. Your Genealogical Family Tree vs. Your Genetic Family Tree
Source: Generated by the author from similar fan charts in "Guide to DNA Testing and Genetic Genealogy" by Blaine T. Bettinger, 2016

On the other hand, the genetic family tree contains only those ancestors who actually contribute to your DNA. Some DNA is lost at each generation due to recombination. We will talk more about this when we discuss autosomal DNA (atDNA) and X-DNA. All of your ancestors 3 - 4 generations back are represented in your DNA, but only about 10% of your DNA remains from further back, for example from your 32 ggg-grandparents and further back. Surprisingly, there is only a 0.01% chance that all 64 of your gggg-grandparents are represented in your DNA. The relevant timeframe for your genetic family tree is 100s-1000s of years ago, but depending on which DNA is lost through recombination, you may or may not match a cousin with a most recent common ancestor 5 or 6 generations back. The genetic family tree is therefore a subset of the genealogical family tree.

This is a good place to end our discussion of the DNA testing procedure. Although this may be more information than you ever wanted to know, it is interesting technology and it helps to provide for a better understanding of the different DNA tests to be discussed in future articles. Next, we will focus on mitochondrial DNA (mtDNA) testing and how this can be used to uncover your genetic ancestry. Until then, happy genealogical and genetic hunting!

