Understanding AncestryDNA Ethnicity Estimations

By Marty Brady

21 November 2020

The first point I want to make is that there is a lot of information on ethnicity estimates and I cannot cover it all in a half hour. If you want to learn more, I refer you to the ethnicity white paper by AncestryDNA, many AncestryDNA articles on the Ancestry website and a 2019 webinar by Mary Eberle on Family Tree Webinars.
https://support.ancestry.com/s/article/AncestryDNA-White-Papers
https://support.ancestry.com/s/article/AncestryDNA-Ethnicity
https://support.ancestry.com/s/article/Getting-the-Most-from-your-AncestryDNA-Test-Results-Part-2-of-3-1460089700555
https://familytreewebinars.com/download.php?webinar_id=945

The basic idea behind ethnicity estimation involves comparing a customer's DNA to the DNA of people with long family histories in a particular region or group (i.e., the reference panel).

AncestryDNA's goal is to have a fast, sophisticated, and accurate method for estimating the historical origins of customer's DNA going back several hundred to over 1,000 years.

The newest approach improves upon the previous version in three ways.

1. The number of possible regions that a customer might be assigned was increased from 61 to 70.
2. The accuracy of the assigned regions was improved.
3. The percentage assigned to each region was improved.

Definitions

Allele – A variant in the DNA sequence. For example, a SNP could have 2 alleles (biallelic), the ancestral allele and the mutational allele. i.e., A or C. In other words, if "A" is the ancestral allele, the customer could have two ancestral alleles (one from Mom and one from Dad). Or if "C" is the mutational allele, the customer could have two C's. Or the customer could have a "C" from Mom and an "A" from Dad or an "A" from Mom and a "C" from Dad. Chemically, the alleles are called nucleotides or bases.

Genotype – A general term for observed variation either for a single site or the whole genome.

Haplotype – AncestryDNA definition - A stretch of DNA along a chromosome.

Haplotype – SNPedia definition - A specific combination of SNPs all occurring together on the same chromosome (i.e. all occurring on the chromosome inherited from Dad, or, inherited from Mom).

SNP - (Single Nucleotide Polymorphism) – A single position (nucleotide) in the genome where different variants (alleles) are seen in different people. Nucleotides are also often referred to as "bases," as in a base pair (bp).

Ploidy – the number of sets of chromosomes (chr) in a cell. Humans are **diploid** organisms, i.e., we have 2 sets of chr (one from Mom & one from Dad).

Diploid - Comes from the Greek "diplous" meaning double + ploidy (defined above).

Haploid - Comes from the Greek "haplous" meaning single + ploidy (defined above). The only cells in humans that are haploid are the sex cells (i.e., gametes, sperms and eggs).

Haplotype - a combination of the terms Haploid + Genotype.

Haplogroup - a grouping of similar **haplotypes**. In genealogy, we talk about 2 types of haplogroups (mitochondrial haplogroups and Y haplogroups) because those are the only two types of DNA in humans that exist as a single chromosome. Men have at most one copy of the Y chromosome (from Dad). In general, humans have one set of mitochondrial DNA (from Mom).

Haplotype is a combination of the terms haploid + genotype. So, it is the composition of the DNA (sequence of SNPs) along a single chromosome (the one we inherited from Mom or the one we inherited from Dad). When they say they are trying to determine the **haplotype** of a DNA segment, they are referring to the sequence of SNPs in a particular window or region on a particular parent's (Mom or Dad) chromosome. Ancestry is able to estimate a customer's ethnicity because certain **haplotypes** are associated with particular regional populations (i.e., Swedish, French, Chinese, etc.)

Steps to Developing the Reference Panel

Step 1 – AncestryDNA selects candidates from published data, the Ancestry customer list and the AncestryDNA proprietary reference collection.

Step 2 - They filter out pieces of DNA between closely related samples.

Step 3 - They use Principal Component Analysis (PCA) to remove discrepant samples (when the pedigree does not agree with the DNA data).

Step 4 - Then the panel is performance tested.

What Samples are Used as the Source of the Reference Panel

Ideally, you would use DNA samples from people who lived hundreds of years ago. But that is not possible/practical.

So, AncestryDNA examined close to 97,000 samples for the Reference Panel.

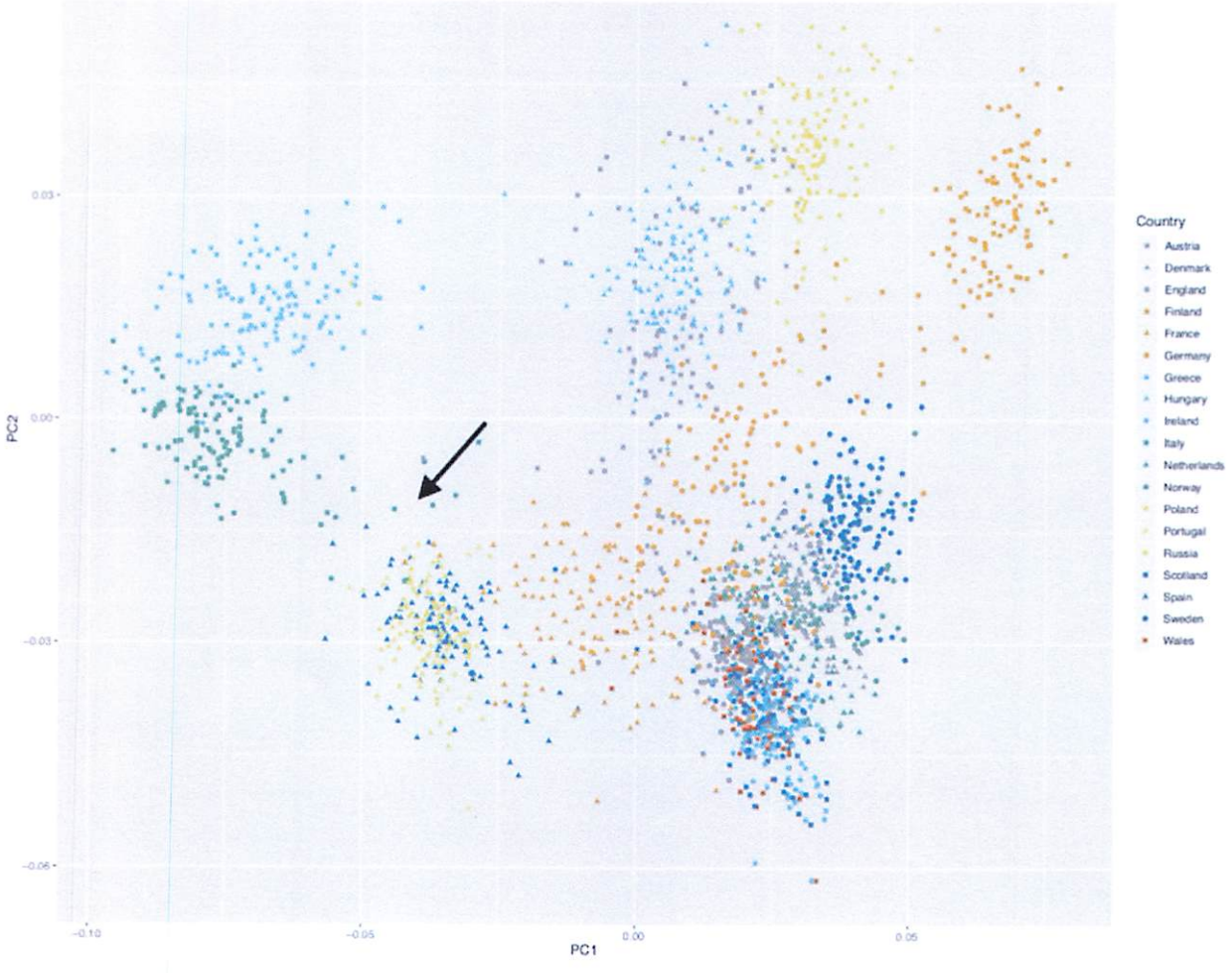
1. Over 1,000 samples from 52 worldwide populations from the Human Genome Diversity Project (HGDP).
2. Over 1,800 samples from 20 populations from the 1000 Genomes Project.
3. Over 900 samples from 91 populations from the Human Origins dataset.
4. Samples from a proprietary AncestryDNA reference collection as well as AncestryDNA samples from customers who had previously consented to research.

Reference Panel Refinement

1. About 300,000 SNPs of the ~700,000 SNPs tested are used for ethnicity estimates.
2. AncestryDNA computes the probability that a particular segment of DNA (an observed **haplotype**) came from for a particular population, for example France or Sweden or China.
3. They then estimate the frequency of this **haplotype** in each population, which requires that people in the reference panel not be closely related which would skew the data. So, AncestryDNA removes candidates that share segments of more than 20cM (IBD).

4. They also remove samples from the reference panel when underlying genetic information disagrees with the pedigree data.
5. After removing PCA outliers, they divide their global reference panel into populations corresponding to distinct genetic clusters in the PCA plots.
6. They remove 5% of samples from the reference panel and estimate their ethnicity using the remaining 95% of samples as the new reference panel. This process is repeated 20X, each time removing a different 5% of the panel and estimating their genetic ethnicities

Example of a PCA plot



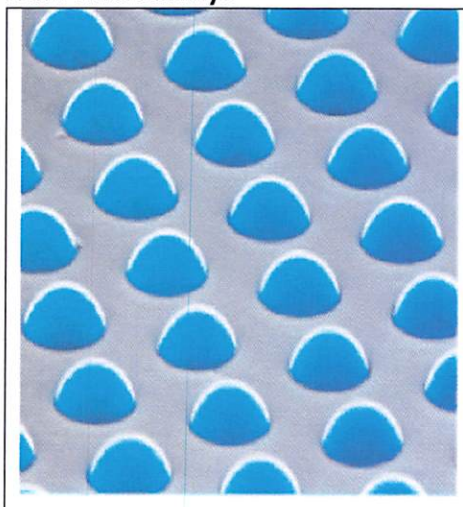
In the above PCA plot, the green squares (about hallway up the left side of the plot) represent people who report an Italian pedigree. The blue and yellow tringles near the arrow represent people who have reported Portuguese and Spanish (P & S) pedigree. The green squares near the arrow are closer to the P & S population cluster than they are to the Italian cluster, so they are removed from the Reference Panel.

How Does AncestryDNA Determine a Customer's Haplotype?

The customer's genome is divided into 1001 windows. Ancestry uses 300,000 SNPs for ethnicity estimates. So, 300K SNPs divided into 1,001 windows yields about 300 SNPs per

window on every chromosome inherited from each parent. AncestryDNA estimates that each window is about 3 to 10 cM.

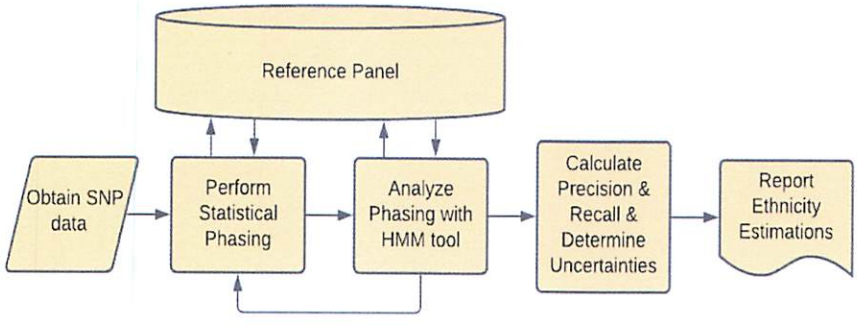
DNA Microarrays



Ethnicity estimates are based on DNA microarray tests. These are glass plates ("chips") with 1.5u beads attached in an array (like a spreadsheet). 700 rows X 1000 columns would yield 700,000 SNPs. For our purposes, you don't need to know how the technology works. Just know that each bead detects a particular SNP. Since our DNA is a mix of Mom's DNA and Dad's DNA, each bead could have 4 possible results. Two ancestral alleles (one from Mom and one from Dad), two mutational alleles, one ancestral allele from Mom and one mutational allele from Dad, or one mutational allele from Mom and one ancestral allele from Dad. It is the goal of AncestryDNA to distinguish those four possibilities at each SNP position.

With the SNP information from a customer's Mom & Dad mixed together on the "chip", it would be difficult to assign a **haplotype** to any segment unless Mom and Dad had the exact same **haplotype**. I liken it to having two decks of cards (decks A & B), shuffling them together, dealing them out and trying to determine what kind a hand you had from deck A (without looking at the back of the cards) and what hand from deck B. So, the very first thing AncestryDNA does is to phase the raw SNP data using statistical algorithms. This means assigning which SNP allele came from Mom and which SNP allele came from Dad.

Flowchart of the Ethnicity Estimate Process



BEAGLE is a commercial software package that includes several algorithms used by AncestryDNA for phasing. AncestryDNA has also written a software package to be run in conjunction with BEAGLE to further refine the phasing results.

The Hidden Markov Model (HMM)

The genetic ethnicity of each position (think SNP) in each window is inferred using a statistical tool called a hidden Markov model (HMM). The customer's genome is divided into 1,001 stretches of DNA called windows & the "hidden" state (ethnicity) is determined for each of them. With the current release, the algorithm has been updated to be able to tell which

ethnicities were inherited from one parent and which ethnicities were inherited from the other parent.

HMMs have two components, called emission & transition probabilities.

Emission probability - Tells how likely it is that a stretch of DNA came from each of the 70 populations based on the observed SNP sequence.

Transition probability - The odds that an ethnicity will change from one window to the next.

Emission Probability

1. Define the Windows.
   Each window covers one section of a single chromosome and is small enough (e.g., 3-10 cM) that both the maternal and paternal **haplotype** in a given window are likely to each come from a single population.
2. Create the **haplotype** models.
   Construct a BEAGLE **haplotype** cluster model for each window using hundreds of thousands of **haplotypes**.
3. Annotate the reference panel.
   Identify the **haplotype** clusters in Ancestry's model that are associated with each population in the reference panel.
4. Compare the test sample to the reference panel to assign population labels using an HMM.
   For each window, both **haplotypes** (Mom's and Dad's) may come from the same population or from different populations, and the resulting emission probabilities are calculated for all possible combinations.

Transition Probability

Because Ancestry's phasing algorithm sometimes misassigns which DNA came from one parent and which DNA came from the other parent, they can use the fact that each parent has different ancestry to improve their assignments of which DNA came from each.

1. It is very unlikely that there would be a change in the same window in both the set of DNA from Mom and the set of DNA from Dad.
2. If Mom is mostly Swedish and Dad is mostly Chinese, and the model encounters a place where it thinks Mom's DNA is Chinese and Dad's is Swedish, it will flip the assignment at that point.

Ancestry runs their HMM on a customer's DNA to find the most likely sequence of ethnicities along the DNA. (i.e., the algorithm takes the "Viterbi" path, the sequence of hidden states (ethnicities) that returns the highest probability).

Proportions are then calculated. For example, a customer with the sequence Sweden/Sweden, Sweden/Sweden, Sweden/Sweden, France/Sweden, France/Sweden would be 20% France and 80% Sweden (if windows have identical size).
To determine the confidence limits, Ancestry randomly samples 1,000 less likely sequences of ethnicities (non-Viterbi paths).

Assessing Precision and Recall

The full definitions are below, but basically **Recall is a measure of the accuracy of the estimate** and **Precision is how narrow a range (in %) is the estimate**. Accuracy and precision are not the same for all ethnicities. AncestryDNA is more accurate and precise in estimating some ethnicities. See the white paper for a listing of the recall and precision for the various reference populations.

Precision – How much of the reported ethnicity is true. For example, if the process predicts an individual has 40% Swedish, but only 30% really is, then the process has a precision of 0.75 for Swedish ethnicity. Precision is expressed as the amount of correctly identified ethnicity divided by the estimated value for that region.

Recall – How much of the true ethnicity is called by the process. If an individual has 50% Swedish ancestry, but the algorithm predicts 40%, the process has a recall of 0.8 for Swedish.

Assessing Uncertainty in Ethnicity Estimates

Ethnicity estimates are not an exact science. So, the uncertainty of the estimate is reported.

If someone is reported as 40% Scotland with a confidence range of 30-60%, this means that they are most likely 40% Scotland but they could be anywhere between 30% and 60% Scotland.

Ancestry's objective when defining this approach was to maximize the probability that the reported range contains the true ancestry proportion (**recall**) while also maximizing **precision** by maintaining a fairly narrow range.

In conclusion

Ethnicity estimations are a complex process sometimes with a lot of uncertainty. Hopefully, some day we will be able to upload our raw DNA data to a website that will phase our data for us and issue us a report of our phased data.