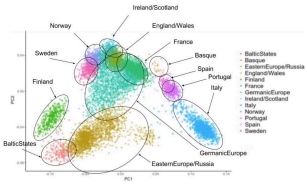


DNA Ethnicity Estimations

by Lee Macklin (not affiliated with AncestryDNA)

Determining where your ancestors lived is not a precise science. Ethnicity estimations are based on mathematical probabilities and comparisons to well documented genealogy histories. This is a two step process. This article is a brief summary of AncestryDNA's "***Ethnicity Estimate 2019 White Paper***".

Step 1

The basic idea behind ethnicity estimation involves comparing a person's DNA to the DNA of people with long family histories in a particular region or group, called a reference panel, and looking for segments of DNA that are most similar. If, for example, a section of a person's DNA looks most similar to DNA in the reference panel from people from Sweden, that section of the person's DNA is said to be from *Sweden*, and so on. The end result is a portrait of a person's DNA made up of percentages of all the ethnicities contained in the reference panel.

Each DNA testing company decides on the number of reference panels as well as the regional breakdown. For example, AncestryDNA's 2018 update contains 40,017 carefully selected people that best represent its 60 global regions. Previously, AncestryDNA had 16,638 people that represented 43 regions. The dramatic increase in the size of their reference panel resulted in more accurate ethnicity predictions and the increased number of regions resulted in more granular global locations. This is also a key reason why your ethnicity estimates change over time with the same DNA testing company as well as across companies (23andMe, MyHeritage, etc.).

This is AncestryDNA's most recent "***Reference Panel***".

Step 2

After establishing and validating a reference panel, the next step is to estimate a person's ethnicity by comparing over 300,000 single nucleotide polymorphisms (SNPs) from their DNA to those of the reference panel. It assumes that an individual's DNA is a mixture of DNA from the 60 populations represented in the reference panel. This is illustrated in Figure 3.1, where, because of DNA recombination, a person inherits long stretches of DNA from his or her four grandparents who, in this example, come from four "single source" reference populations.

"Because DNA is passed down from one generation to the next in long segments, it is likely that the DNA at two nearby locations in the genome were inherited from the same person and so the same population. This means we can get more accurate results by looking at multiple nearby SNPs together as a group, or "haplotype", instead of looking at each SNP in isolation."

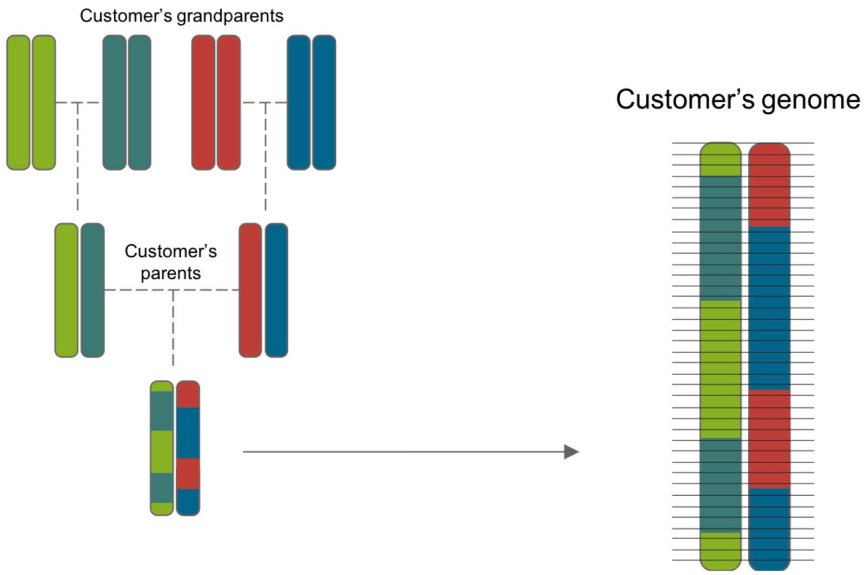


Figure 3.1 - ***Inheritance of DNA from different populations.*** *On the left, we present a three-generation genetic family tree. For each individual, we show two vertical bars representing the two copies of a single chromosome present in each individual. These bars are colored by the reference population from which they inherited their DNA. Each of the four grandparents (solid bars, top row) has inherited 100% of their DNA from a single population that is different from the other three. The DNA is passed forward to the parents and finally to the customer, who, through the process of recombination and assortment, ends up inheriting a shuffled set of chromosomes from each parent. The colors show that the customer's DNA is a mixture of the DNA inherited from their four grandparents, with long stretches inherited from the same grandparent. On the right, we show that to obtain a customer's ethnicity estimate, we divide the customer's genome into small windows (represented by black horizontal lines). For each window we assign a single population to the DNA within that window inherited from each parent, one population for each parental haplotype. Each window gets a population assignment based on how well it matches genomes in the reference panel.*

"We estimate a person's genetic ethnicity by assuming that each segment of their genome comes from one of the 60 populations in the reference panel. We divide the person's genome into 1,001 windows. We assume that each window is small enough that each of the two parental haplotypes present in the window came from exactly one population. We then combine information from all the windows to estimate what overall portion of the person's genome came from each of the populations."

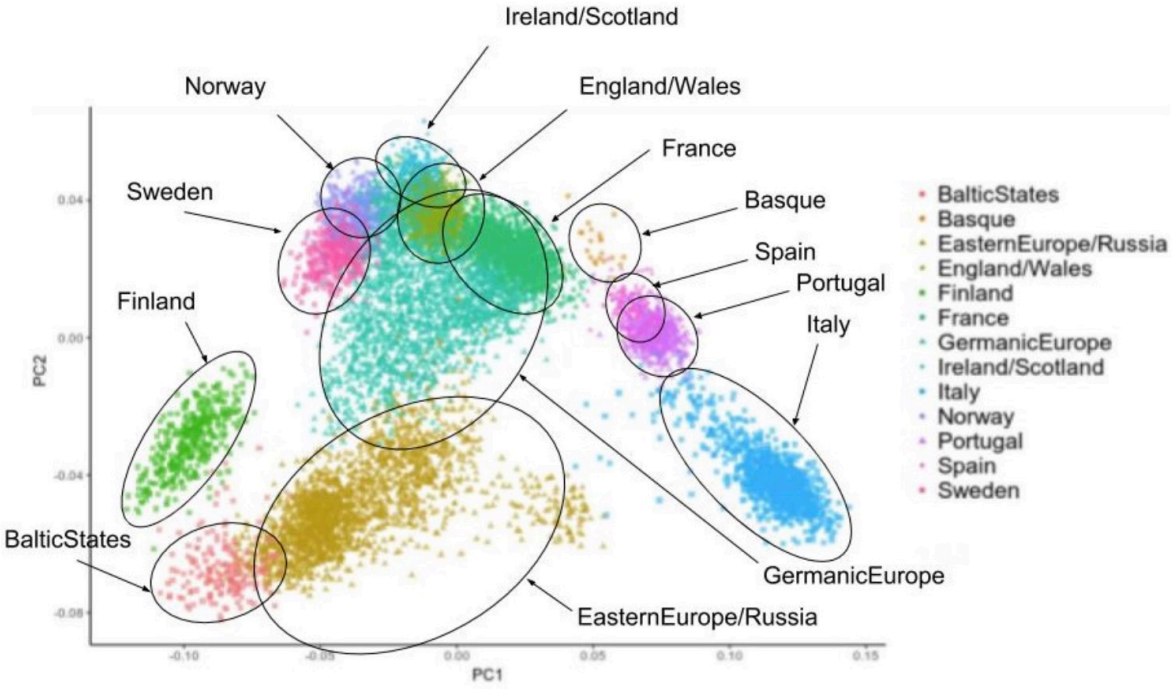
As you can see in Figure 3.1, each window does not need to have a single ethnicity associated with it. Instead, it can have one from one parent and one from the other. For example, the first window has two different ethnicities represented by the colors green and red. Any ethnicity estimator that uses the technology AncestryDNA does to read DNA has to account for the possibility of two separate ethnicities in each window. In other words, it has to employ a model that looks at the DNA and can identify the DNA

as a mix of red and green as opposed to just red or just green (or any of the other possible combinations).

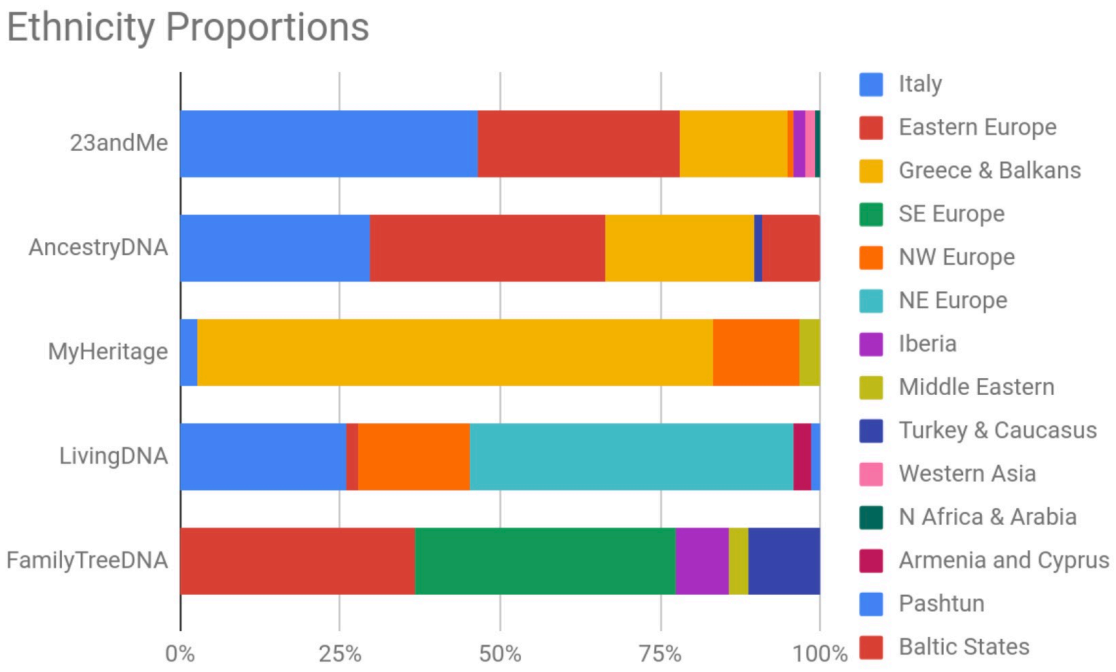
This is a sensible model for human DNA because human genomes are organized linearly along chromosomes. Additionally, the nature of inheritance means that whole segments of the genome and, therefore, many consecutive nucleotides that AncestryDNA looks at along a chromosome, will have the same DNA ancestry.

“AncestryDNA uses microarrays to obtain DNA data from customer samples. We look at over 700,000 individual locations on the DNA (SNPs) and determine the nucleotides at each position. For example, we may see an A and a T at position 1, a G and a G at position 2, and so on. We use around 300,000 of these SNPs in the ethnicity estimate.”

“In working with data from arrays, it is important to remember that people have two copies of each of the 22 chromosomes that AncestryDNA reports data back on. One set of chromosomes comes from Mom and the other from Dad. This means there are two results for each position AncestryDNA analyzes, and those results must be interpreted to assign which DNA came from which set of chromosomes (this process is called phasing). AncestryDNA must consider what possible combinations of ethnicities might look like. For example, if one customer has a section of their DNA that came from Swedish ancestors from Mom’s side of the family and Japanese ancestors on Dad’s, the algorithm must be able to distinguish this from a second customer with Swedish and Nigerian ancestors.”



AncestryDNA’s Principle Component Analysis Plot



A Person's DNA Ethnicity Results Comparison - 23andMe and AncestryDNA are Close

This article explains how DNA Ethnicity estimates are determined. Since DNA analysis has been around for about 30 years now, reference panels have been created and continue to be refined. That is why every couple of years the DNA testing companies revise their predictions as they factor in the results of more and more reference panels. Today, the shared DNA amounts and accompanying relationship predictions are pretty accurate. But as the above diagram illustrates, there are significant differences among the DNA testing companies. This is not because the data is inaccurate, rather it is the result of how they chose to group and compare your genetic data.