Ethnicity Estimations

By Martin Brady, 21 May 2022

Introduction

What is the difference between ethnicity and historical origin? The Ancestry white paper entitled "Ethnicity Estimate 2021 White Paper" refers to their process as estimating the "historical origins" of customers. Ancestry uses DNA for their estimates. DNA does not give you information about the traditions, language or culture of a population. It can provide information about the geographical origins of your ancestors. So, it is probably more accurate to call this historical origin estimates instead of ethnicity estimate. That being said, since everyone refers to this process as ethnicity estimation, I will usually refer to this process as ethnicity estimation. In the Paul Woodbury webinar referenced below he gives some observations about ethnicity as listed below.

> "Ethnicity is a grouping who identify with each other based on shared attributes that distinguish them from other groups such as traditions, ancestry, language, culture, history or religion. Because of this, individuals of the same ethnicity often belong to the same population, and in turn may share a similar gene pool. However, it is important to keep in mind that ethnicity is much more than the shared DNA of people in a particular population. While DNA may be associated with a particular ethnicity, it is not, in and of itself, the sole defining feature of ethnicity. Therefore, it is possible for an individual to have ethnic admixture from a particular region and not be a part of that ethnicity just as it is possible for an individual to be a part of an ethnicity and yet have no genetic association with that ethnicity."

With that caveat in mind, the main way for the core five genetic genealogy (GG) companies to ascertain the historical origins of customers is by examining their DNA. Most companies examine around 700,000 genetic markers known as SNPs (Single Nucleotide Polymorphisms) for each customer. Ancestry uses about 300,000 of these SNPs for historical origin estimates. (The other 400,000 SNPs are used for things such as health and traits information). The 300,000 SNPs of each customer is compared to a reference population to determine what population the customer's SNPs most closely resemble. Knowing how humans have historically migrated over whole continents and from continent to continent how is this comparison even possible? The answer is that humans in general over the last 1000 years have in general not migrated as far and wide as they did 20,000 years ago. At least, that is the theory and, as always, there are exceptions.

If all customers' ancestors were isolated on remote islands with no contact with other populations, determining the ethnicity of customers' ancestors would be easier. Things that can change our expected ethnicity estimates are conquests such as the Mongol or Roman empire. The Mongol empire in 1259 extended from about 60 miles east of Venice to Japan. That is a distance of over 6000 miles. The northern and southern edges of the Mongol empire

included Moscow and India. From the facial features of Vladimir Putin, I would be surprised if he didn't have Mongol heritage. The Lombards, a Germanic people, ruled Northern Italy for hundreds of years. I have a friend who didn't believe the Ancestry results that said she had German DNA because she was born in Italy (just north of Venice). Even though there were conquests and movement of individuals from one area to another, most recent ancestors remained in the same place for long periods of time and the five genetic genealogy companies have been reasonably successful in determining customers' historical origins by examining SNP markers. How do the core 5 GG companies do that?

I highly recommend Diahan Southard's video on Family Tree Webinars for a very simple explanation of the process. Her website also has a lot more information on ethnicity estimates. For an in-depth review of the process, AncestryDNA provides an excellent white paper on the subject. Since Ancestry provides the most information on the process (followed by 23andMe), I will be using their paper extensively to explain the process.

The Reference Panel

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). So, Ancestry must first establish a reference panel. Ideally, using DNA samples from people who lived hundreds of years ago (which is not usually possible).

They collect reference samples from different reference populations and filter out closely related individuals and those whose DNA differs from their pedigree. They then run multiple tests to see how well their reference population can identify samples of known genetic origin. Ancestry is constantly refining the process. In 2017, Ancestry used about 45,000 reference samples to define 70 reference populations. In 2021, Ancestry used almost 57,000 reference samples to define 77 reference populations. As the process is refined, the historical origins estimates become more granular, more precise.

The Need to Phase

Haplotypes (see definition below) can only be determined when you know which SNP values (bases) belong on which copy of a particular chromosome. Since the GG company does not know which SNP value came from which parent, they must first assign the appropriate base to the appropriate copy of the chromosome. This is called phasing. The GG company uses statistics to phase the data. This is a complex process and can result in "misassignment" of a base to the wrong copy of the chromosome. If we look at 4 of my SNPs from chromosome 1 (see below), we can see that there are at least 8 different ways to assign the bases to the 2 copies of chromosome 1. Since Ancestry must phase 300,000 SNP alleles, you can see that there are a lot of possible ways to assign those SNPs and a lot of chances for misassignment.

SNP id	chromosome	position	Allele 1	Allele 2
rs3131972	1	752721	A	G
rs114525117	1	7599036	A	G
rs4040617	1	779332	A	G
rs4422948	1	835499	A	G

Possible assignments include the following:

#	Assignment		#	Assignment
1	AAAA and GGGG		5	AAGG and GGAA
2	AAAG and GGGA		6	AGGA and GAAG
3	AAGA and GGAG		7	AGAG and GAGA
4	AGAA and GAGG		8	AGGG and GAAA

So, statistical phasing adds uncertainty to the process from the outset. Also, misassignments could have a greater effect on small segments vs large segments. In the future, it may be possible to analyze each copy of each chromosome from end to end in one long sequence so they won't have to "phase" the SNPs. They will know from the data which bases belong on the same copy of a chromosome. But for now, statistical phasing is necessary.

Haplotype Comparisons

When the big 5 GG companies analyze a customer's DNA, they determine the customer's group of SNP markers on both copies of each chromosome. This is known as the customer's haplotype.

- The haplotype of a DNA segment refers to the group of SNPs in a particular window or region on the copy of the chromosome inherited from a particular parent (Mom or Dad).
- Certain haplotypes are associated with particular populations.

When Ancestry plots a particular set of SNPs against another set of SNPs similar populations tend to cluster on the plot. Ancestry uses this information to help assess the accuracy of their reference population. Individuals that identify with a particular ethnicity but who don't cluster with that ethnicity's set of SNPs are eliminated from the reference population.

Because recombination creates segments of DNA that may be associated with different ancestors and therefore different historical origins, you can't compare the haplotype of an entire copy of an autosomal chromosome to that of the same chromosome in the reference population. It would not yield sensible data. For example, if a copy of chromosome 1 was a mixture of segments from French and Chinese grandparents, the haplotype of the entire copy of the chromosome would be something like Frinese, which wouldn't make sense. So, the GG companies divide the genome into 1001 windows and compare the haplotype of each window of each copy of each chromosome to the haplotype of the corresponding window in the reference population. If they test 300,000 SNPs and they have about 1000 windows, that means there will be on average about 300 SNPs in each window.

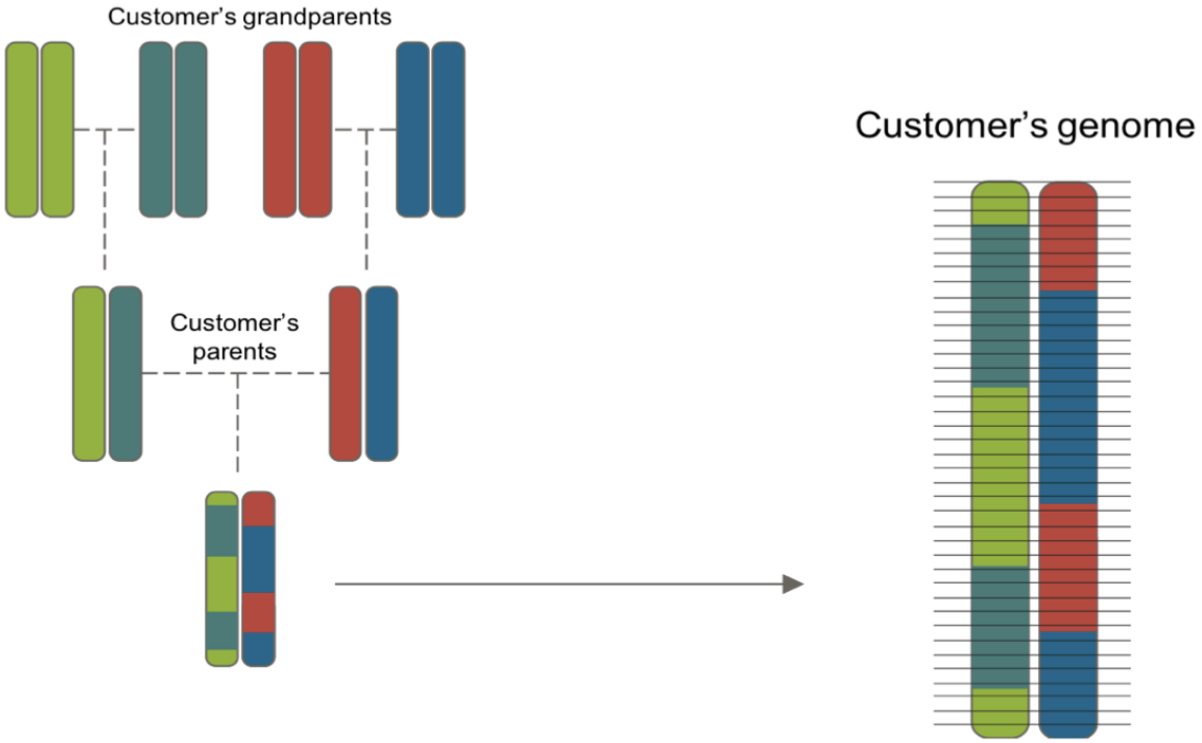

Figure 3.1: Inheritance of DNA from different populations. *On the left, we present a three-generation genetic family tree. For*

(the above image is from The Ancestry white paper on Ethnicity Estimates)

Since Ancestry knows that each ethnicity segment on the chromosome will probably contain many windows, the ethnicity of successive windows will not change very often. In fact, a change would only be expected at recombination points (if then). If the statistical algorithm encounters flip-flopping, it can adjust the phasing. If it knows one parent is Chinese and another is French and it finds that these determinations have flip-flopped from one copy of the chromosome to the other, the phasing software can simply reverse these sections.

Precision and Recall

Ancestry assesses how well it has called the historical origins of the reference population using precision and recall assessments. The more accurate the call the closer the Precision and Recall values will be to 1. A Recall value of 0.53 for France means that the algorithm only calls about 53% of the true ethnicity for that population. Ancestry does not call each population with the same accuracy. A table of **some** of the reference population values is below.

Precision – How much of the reported ethnicity is true. For example, if the process predicts an individual has 40% Chinese, but only 30% really is, then the process has a precision of 0.75 for Chinese 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% Chinese ancestry, but the algorithm predicts 40%, the process has a recall of 0.8 for Chinese.

Region	Precision	Recall
Benin & Togo	0.85	0.96
England & Northwestern Europe	0.52	0.7
European Jewish	0.92	0.99
France	0.84	0.53
Germanic Europe	0.8	0.61
Indigenous Cuba	0.76	0.17
Ireland	0.54	0.94
Melanesia	0.98	0.99
Norway	0.63	0.89
Southern China	0.88	0.92
Southern Italy	0.83	0.61
Spain	0.8	0.63

Uncertainty

In science, there is always uncertainty. Results are usually expressed in ranges to quantify that uncertainty. Ethnicity estimates are not an exact science. If someone is reported as 40% French with a confidence range of 30-60%, this means that they are most likely 40% French but they could be anywhere between 30% and 60% French. When an estimate is given as a low percentage, the range usually includes 0%, which means that the customer may not have that ethnicity at all. In addition, the closer two populations are, the greater the uncertainty in the call is and therefore the wider the range will be.

You can examine your results on 23andMe using 50% to 90% confidence levels and see if it changes your estimate.

Chromosome Painting

You can look at your ethnicity chromosome painting on 23andMe. My chromosome painting is so uniform, I could paint it with a roller. But many people have diverse ethnicities and can discover a lot of information about the ethnic segments by examining their ethnic chromosome painting. For example, if two unusual ethnicities seem to be immediately adjacent on several chromosomes, it could be that they were inherited together from one parent.

The size of the ethnic segments could be an indication of how long ago they were inherited. 23andMe offers a feature called Your Ancestry Timeline that shows you how far back an ethnicity of interest goes. They have a white paper on the subject which I haven't read yet, but you may find it illuminating.

Side View
Ancestry has a new feature that estimates which ethnic segments you received from which parent. It works by comparing bits of DNA of people from your match list with your DNA. The assumption is that most matches will only match you on either your father's or your mother's side. When they determine which parent a particular section of DNA matches, they know that the corresponding section on the other copy of the chromosome matches the other parent. The more sections they compare, the more accurate they can be with the assignment.

Communities
Communities are calculated from the tree information of our DNA matches. The Ancestry communities in 2017 totaled 1100+ regions. In 2021 they total 1500+ regions. So, this is an area that is greatly improving and it provides unbelievably accurate granular estimates of the communities our ancestors belonged to. The feature is also on 23andMe but not as granular. Ancestry's white paper on this subject is illuminating for those who are interested.

MyHeritage feature
MyHeritage allows you to filter your matches based on ethnicities, locations or genetic groups. This can be quite useful if you are looking to trace an unusual ethnicity.

Conclusion
There is a lot of information that we can derive from ethnicity estimates and the ethnicity associated tools provided by the 5 GG companies. Also, the big 5 GG companies keep improving their processes, so it is valuable to keep up with their new information. However, historical origin estimates currently have a lot of uncertainty and they are the most valuable when considered along with documentary evidence.

References

Eupedia website
https://www.eupedia.com/europe/Haplogroup_R1b_Y-DNA.shtml#migration_map (Accessed 20 May 2022)

https://www.ancestrycdn.com/support/us/2021/09/ethnicity2021whitepaper.pdf (Accessed May 17, 2022).

Diahan Southard webinar April 2021 (Accessed May 17, 2022).
https://familytreewebinars.com/webinar/four-factors-influencing-your-dna-ethnicity-results/

Your DNA Guide website – Explanation and free downloadable guide (Accessed May 17, 2022)
https://www.yourdnaguide.com/ethnicity-estimate

Where Did That Come From?! Tracing the Origins of Unique Ethnicity Admixture by Paul Woodbury 24 September 2021 (Accessed May 17, 2022)
https://familytreewebinars.com/webinar/where-did-that-come-from-tracing-the-origins-of-unique-ethnicity-admixture/?search=woodbury

The Stories Behind the Segments by Blaine Bettinger 1 Oct 2019 (Accessed May 17, 2022)
https://familytreewebinars.com/webinar/the-stories-behind-the-segments/

If you want more info on Principle Component Analysis
Stat Quest PCA video
https://www.youtube.com/watch?v=HMOI_lkzW08 (Accessed 20 May 2022)

https://www.youtube.com/watch?v=_UVHneBUBW0 (Accessed 20 May 2022)