Ethnicity Estimate 2023 White Paper

Jeffrey Adrion, Jenna Lang, Keith Noto, Richard Olpin, Alisa Sedghifar, Yong Wang, Aaron Wolf (in alphabetical order)

Summary:

The AncestryDNA® science team has developed a fast, sophisticated, and accurate method for estimating the historical origins of customers' DNA going back several hundred to over 1,000 years. Our newest approach improves upon our previous version in the number of possible regions that a customer might be assigned (from 84 to 88) as well as an increase in accuracy to both regions assigned and the percentages assigned for each region. We have added four new regions as well as made improvements to the composition of our reference panel, resulting in more accurate estimates overall. Given the cutting-edge nature of this type of science, we will continue to refine our approach and improve estimates.

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, what we call a reference panel, and looking for segments of DNA that are most similar. If, for example, a section of a customer's DNA looks most similar to DNA in the reference panel from people from Norway, that section of the customer's DNA is said to be from Norway, and so on. The end result is a portrait of a customer's DNA made up of percentages of the 88 regions contained in the reference panel.

That is a short version of how AncestryDNA determines a customer's ethnicity estimate. The rest of the white paper will delve more deeply into:

1. How the reference panel samples are chosen, their makeup, and how the panel is validated
2. How the algorithm that determines a customer's genetic ethnicity works and how it is validated

1. Introduction

Genetic ethnicity estimates that determine which populations in a reference panel are most similar to someone's DNA are a major component of the Origins product provided by AncestryDNA. As its name suggests, Origins provides customers with insights into their past by analyzing their DNA.

AncestryDNA has employed a team of highly trained scientists with backgrounds in population genetics, statistics, machine learning, and computational biology to develop a fast, sophisticated, and accurate method for estimating genetic ethnicity for our customers. In this document, we describe the approach we use to make those estimates. We will discuss the development of the reference panel we compare each customer sample against, the inference method we apply to estimate genetic ethnicity, and finally the extensive testing regimen we employ to assess the quality of our estimates.

Glossary

Admixed — Having ancestry from multiple populations.
Allele — A variant in the DNA sequence. For example, a SNP (defined below) could have two alleles: A or C.
Centimorgan (cM) — A unit of genetic length in the genome. Two genomic positions that are a centimorgan apart have a 1% chance during each meiosis (the cell division that creates egg cells or sperm) of experiencing a recombination event between them.
Chromosome — A large, inherited piece of DNA. Humans typically have 23 pairs of chromosomes with one copy of each pair inherited from each parent.
Genome — All of someone's genetic information; the DNA on all chromosomes.
Genotype — A general term for observed genetic variation either for a single site or the whole genome.
Haplotype — A stretch of DNA along a chromosome.
Hidden Markov model (HMM) — A statistical model for determining a series of hidden states based on a set of observations.
Locus — A location in the genome. It could be a single site or a larger stretch of DNA.
Microarray — a DNA microarray is a way to analyze hundreds of thousands of DNA markers all at once.

Nucleotide — DNA is composed of strings of molecules called nucleotides (also called bases). There are four different types, and they are usually represented by their initials: A, C, G, T.
Population — A group of people.
Phasing — The assignment of DNA to contiguous segments corresponding to the DNA inherited from Mom or Dad. This is done with an algorithm.
Recombination — Before chromosomes are passed down from parent to child, each pair of chromosomes usually exchange long segments between one another and then are reattached in a process called recombination.
Single nucleotide polymorphism (SNP) — A single position (nucleotide) in the genome where different variants (alleles) are seen in different people.

2. Reference Panel

2.1 Calculating an Ethnicity Estimate

Two chromosomes from the same geographic region or the same population will share more DNA with one another than will two chromosomes from different regions or groups. So two pieces of DNA with a historical connection to Portugal will have more DNA in common than will a piece of DNA from Korea and a piece of DNA from Portugal. This is the basic premise behind the ethnicity estimate AncestryDNA provides to its members.

To create the ethnicity estimate, we compare a customer's DNA to a panel of DNA from people with known origins (referred to as the reference panel) and look to see which parts of the customer's DNA are similar to those from people represented in groups in the reference panel. If, for example, a section of a customer's DNA is most similar to the reference panel samples from Senegal, then we identify that section of the customer's DNA as coming from Senegal.

The accuracy of our ethnicity estimate depends on the quality of our reference panel. Because of this, AncestryDNA has invested a significant amount of effort in developing the best possible set of reference samples.

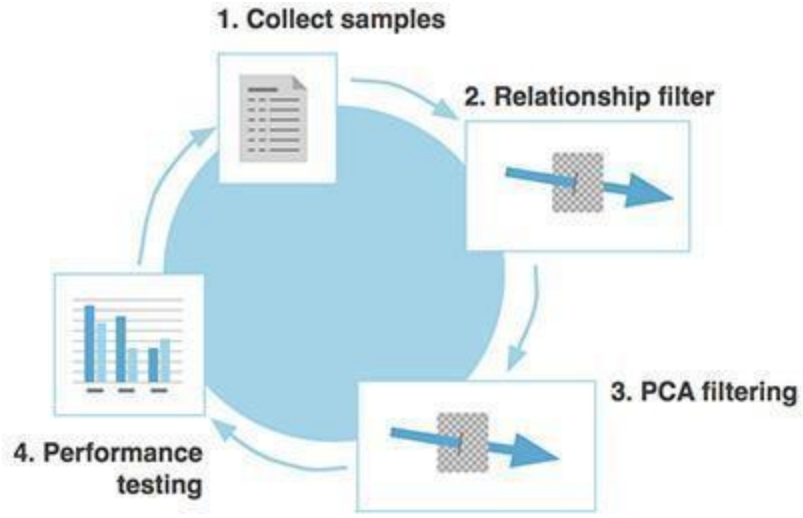


Figure 2.1: Reference Panel Refinement Cycle. *Schematic of the ethnicity estimation reference panel refinement cycle. In* ***step 1*** *we select candidate reference samples from published data, the AncestryDNA customer list, and the AncestryDNA proprietary reference collection. For AncestryDNA samples we rely on pedigree data to select those with deep ancestry from a single population. In* ***step 2*** *we filter out pieces of DNA between closely related samples from the candidate list. In* ***step 3*** *we use principal component analysis (PCA) to remove samples that show a disagreement in pedigree and genetic origin. We also use PCA to guide the identification of population groups. In* ***step 4*** *the panel is performance tested using numerous metrics and compared to the previous release. The final result is a high-quality, well-tested reference panel. The entire procedure is cyclic, and AncestryDNA will continue to make improvements to the panel with the goal of providing the most accurate ethnicity estimation possible with the data available.*

The rest of Section 2 describes the steps taken to develop our current reference panel, including sample selection, quality control, and testing. The ethnicity update that we describe here is not only an update of the reference panel from our 2022 version but also increases the number of global regions from 84 to 88.

2.2 Who should be included in the reference panel?

Identifying the best candidates for the reference panel is key to providing the most accurate ethnicity estimate possible from a customer's DNA sample. Under perfect circumstances, we would construct our reference panel using DNA samples from people who lived hundreds of years ago. Unfortunately, it is not yet possible to reliably sample historical populations in this way. Instead, we must rely on DNA samples collected from people alive today and focus on those who can trace their ancestry to a single geographic location or population group.

When asked to trace familial origins, most people can only reliably go back one to five generations, making it difficult to find individuals with knowledge about more distant ancestry. This is because as we go back in time, historical records become sparse, and the number of ancestors we have to follow doubles with each generation.

Fortunately, knowing where someone's recent ancestors were born is often a sufficient proxy for much deeper ancestry. In the recent past, it was much more difficult and thus less common for people to migrate long distances. Because of this, the birthplace of a person's recent ancestors often represents the location of that person's deeper ancestral DNA.

AncestryDNA Reference Panel Candidates

In developing the most recent AncestryDNA ethnicity reference panel, we began with a candidate set of close to 180,000 samples. First, we examined over 1,000 samples from 49 worldwide populations from a public project called the Human Genome Diversity Project (HGDP) (Cann *et al*. 2002; Cavalli-Sforza 2005), over 2,500 samples from 19 populations from the 1000 Genomes Project (McVean *et al*., 2012), and over 900 samples from 84 populations from the Human Origins dataset (Lazaridis et al. Nature 2014). Second, we examined samples from a proprietary AncestryDNA reference collection as well as AncestryDNA samples from customers who had previously consented to research. Most of the candidates were selected from the last two groups only after their family trees confirmed that they had a long family history in a particular region or within a particular group. A small number of candidates were selected without a deep family tree, but these passed the rigorous vetting process outlined below. Although it was not possible to confirm family trees for HGDP, Human Origins, and 1000 Genomes Project samples, these datasets were explicitly designed to sample a large set of distinct population groups representing a global picture of human genetic variation.

Reference Panel Candidates from Admixed Populations

In some parts of the world, indigenous people carry DNA from populations originating from continents other than their own. For example, people of Amerindian descent in North and South America may also have some ancestry from Europe and Africa. When creating reference panel regions reflecting geographic regions for the Americas and Oceania, we wanted to use only the parts of the genome with ancestry from the indigenous populations. We did this by looking at our previous ethnicity assignments and choosing only the segments of DNA (or windows) where both chromosomes had assignment to an

ethnicity region corresponding to the indigenous population. So, whereas most of our regions use DNA from the entire genome of each reference panel candidate, when creating reference panels for populations in areas that are now admixed we only use a fraction of each person's genomes. The ethnicity regions where we employ this approach are:

Indigenous Americas–Bolivia & Peru
Indigenous Americas–Colombia & Venezuela
Indigenous Americas–Mexico
Indigenous Americas–North
Indigenous Americas–Yucatan Peninsula
Indigenous Americas—Central
Indigenous Americas—Chile
Indigenous Americas—Ecuador
Indigenous Americas—Panama & Costa Rica
Indigenous Eastern South America
Indigenous Puerto Rico
New Zealand Maori
Aboriginal & Torres Strait Islander
Guam
Hawaii
Samoa
Tonga

For two other regions, Indigenous Cuba and Indigenous Haiti & Dominican Republic, we used windows where only one chromosome had assignment to an ethnicity region corresponding to the indigenous population. We then combined single chromosomes from two different people in the same window. We did this to create a window homozygous for the indigenous region assignment.

2.3 Reference Panel Quality Control

For each sample, we analyzed a set of approximately 300,000 SNPs that are shared between the Illumina OmniExpress platform and the Illumina HumanHap 650Y platform, which was used to genotype HGDP samples. After samples with large amounts of missing data were removed, we filtered out those which were likely to degrade the performance of the reference panel. Samples were typically removed because

they were closely related to another reference sample or the underlying genetic information about a sample's origins disagreed with the family tree data.

When we perform genetic ethnicity estimation, we are interested in computing the probability that a particular segment of DNA, an observed haplotype, is most similar to a region in the reference panel (see Section 4 below). The regions in the reference panel can provide clues as to where someone's ancestors are from, although they do not always directly indicate a person's ancestral origins.

To compute the probability a haplotype is most similar to a reference panel population, we need to estimate the frequency of this haplotype in each population. This requires that people in the reference panel not be closely related. DNA segments shared as a result of recent ancestry, as identified through identity by descent (IBD), do not represent independent haplotypes in a population. Retaining these shared segments can distort the estimates of haplotype frequencies in a population. To avoid this bias, we remove shared segments for candidates that share more than 20 cM of IBD DNA. Details about our approach for detecting shared segments of IBD DNA can be found in our AncestryDNA Matching White Paper.

Next, we remove samples from the reference panel candidate set when the genetic data about ethnicity disagrees with what that person has reported about their ethnicity—when underlying genetic information disagrees with the pedigree data. We identify these outliers using two approaches: (1) we identify clear outliers using our previous ethnicity estimate version, and (2) we use principal component analysis (PCA). PCA is frequently used for exploratory data analysis in population genetics research (Jackson 2003). When applied correctly to genotype data, PCA can capture the genetic variation separating distinct populations (Patterson 2006).

We apply PCA to the samples that have made it through the previous screening processes and plot the early stages of the analysis, the "first four principal components," as a series of scatter plots. We color each sample by its country of origin, determined by pedigree for Ancestry samples and by sample label for public samples (see Figure 2.2).

Figure 2.2: PCA Analysis on European Panel Candidates. *Scatter plot of the first two components from a principal component analysis (PCA) of candidate European samples for the AncestryDNA reference panel. Visual inspection of PCA is useful for numerous aspects of data QC. First, it can be used to identify individual outliers, such as the Italian samples (green squares) that appear near the Portugal and Spain (yellow and blue triangles, respectively) cluster. It can also be useful for identifying poor sample grouping. Finally, it can reveal regions where there is limited genetic separation and clusters overlap (e.g., England, Ireland, Wales, and Scotland clusters) and regions that can be further subdivided.*

Each population tends to form a cluster of points (each point is a sample) in the scatter plot. This is because points that are more genetically similar are closer in PCA space. Helpfully, these clusters of points tend to match geography as well because most people are genetically more similar to others from nearby. Furthermore, these plots quickly reveal outlier samples that are not near other samples from the same population. For example, the green squares near the yellow and blue triangles indicate samples with family trees from Italy whose DNA is more similar to people from Portugal and Spain. These are

examples where the specified population of origin disagrees with the genetic origin represented in PCA space. We also use a related approach called UMAP (Uniform Manifold Approximation and Projection) to visualize population structure and identify outliers (Diaz-Papkovich 2019).

We visually inspect candidates for removal based on a scatterplot like the one in Figure 2.2. Because different collections of samples reveal different amounts of population structure, PCA and outlier removal are repeated for different subsets of data. We first remove outliers at the global level (all samples together), then at the continental level (e.g., outliers in a PCA using only European samples), then at the regional level (e.g., outliers in a PCA of all Scandinavian samples), and finally at the population level (e.g., outliers from a PCA of Norway).

2.4 Iterative Reference Panel Refinement

After removing PCA outliers, we divide our global reference panel into populations corresponding to distinct genetic clusters in the PCA plots. Before using the reference set to estimate ethnicities of AncestryDNA customers, we first determine its quality by measuring the performance of our ethnicity estimation on the reference set itself. This tells us how our ethnicity estimation does on samples that by definition are 100% of a single ethnicity.

To do this, we employ a cross-validation approach. Specifically, we remove 5% of samples from the reference panel and estimate their ethnicity using the remaining 95% of samples as the new reference panel. We repeat this process 20 times, each time removing a different 5% of the panel and estimating their genetic ethnicities using the remaining 95%. We then look at the average predicted ethnicity for samples from each region in the reference set using the results of these cross-validation experiments. Figure 2.3 shows the results of this experiment as box plots.

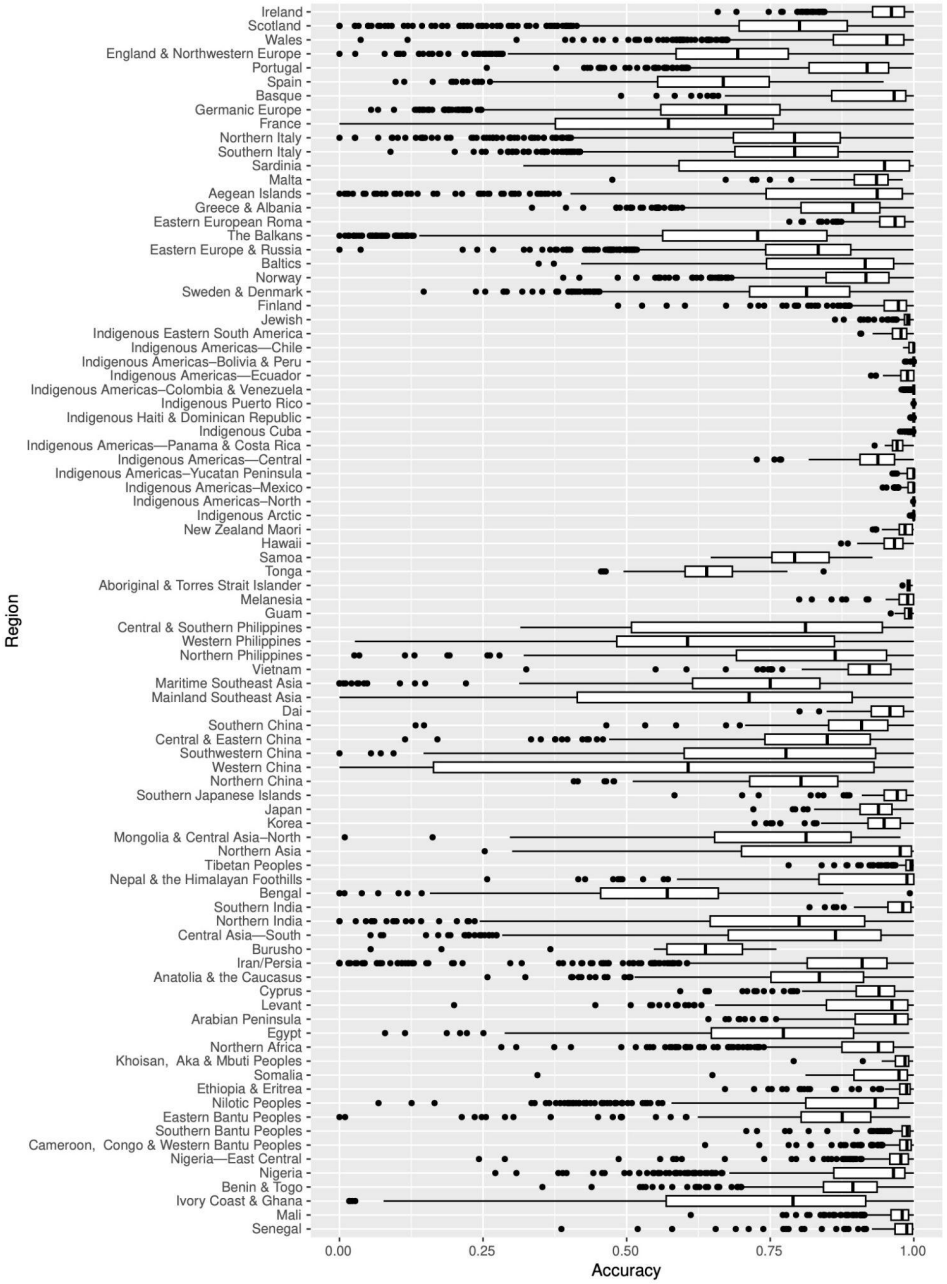


Figure 2.3: 20-fold cross-validation analysis of the Ethnicity 2023 reference panel. *Here we plot the results of an experiment in which 5% of samples are removed from the reference panel, and their ethnicity is estimated using the remaining (95%) panel samples. Each boxplot represents the distribution of estimated ethnicity for all samples from a given region (75%, 50%, and 25% percentiles of estimated ethnicity). 76% of our regions have an average accuracy of 76% or greater. However, there are exceptions. In particular, our average prediction accuracy for samples Burusho and Bengal are not quite as high. There are many factors affecting the accuracy of these numbers, most importantly the number of reference samples in the panel for each region and the genetic distinctness of each region.*

The purpose of this analysis is twofold. First, reference panel samples with extremely poor performance in the cross-validation analysis are removed, as they may poorly represent their ethnic group of origin. Second, the cross-validation experiments allow us to demonstrate our ability to accurately estimate the ethnicities of our reference panel samples using our ethnicity estimation method (see section 3) and thus help us redefine population boundaries. For example, we may merge two populations if performance in the cross-validation experiment is poor in each group but is found to be better in a merged group.

After performing several rounds of reference panel refinement based on cross-validation experiments, we settled on dividing our latest reference panel into 88 global regions. These regions are described in further detail below.

2.5 Updated Reference Panel

The updated AncestryDNA ethnicity estimation reference panel contains 71,306 samples carefully selected as described above to represent 88 global regions (Table 2.1), each with a unique genetic profile. As a comparison, our previous panel of 68,714 samples represented 84 distinct global regions.

Table 2.1: The Final AncestryDNA 2023 Ethnicity Reference Panel

Region	Number of Samples
Aboriginal & Torres Strait Islander	54
Aegean Islands	1111
Anatolia & the Caucasus	426
Arabian Peninsula	208
Baltics	237
Basque	118
Bengal	291
Benin & Togo	319
Burusho	25
Cameroon, Congo & Western Bantu Peoples	502
Central & Eastern China	359
Central & Southern Philippines	300
Central Asia—South	622

Cyprus	282
Dai	90
Eastern Bantu Peoples	179
Eastern Europe & Russia	1709
Eastern European Roma	300
Egypt	396
England & Northwestern Europe	2291
Ethiopia & Eritrea	131
Finland	366
France	5305
Germanic Europe	3408
Greece & Albania	638
Guam	157
Hawaii	363
Indigenous Americas–Bolivia & Peru	269
Indigenous Americas–Colombia & Venezuela	3117
Indigenous Americas–Mexico	581
Indigenous Americas–North	1985
Indigenous Americas–Yucatan Peninsula	316
Indigenous Americas—Central	2076
Indigenous Americas—Chile	539
Indigenous Americas—Ecuador	662
Indigenous Americas—Panama & Costa Rica	466
Indigenous Arctic	36
Indigenous Cuba	9559
Indigenous Eastern South America	2671
Indigenous Haiti & Dominican Republic	1994
Indigenous Puerto Rico	3601
Iran/Persia	1255
Ireland	755
Ivory Coast & Ghana	315
Japan	202
Jewish	447

Khoisan, Aka & Mbuti Peoples	59
Korea	294
Levant	271
Mainland Southeast Asia	516
Mali	504
Malta	97
Maritime Southeast Asia	107
Melanesia	66
Mongolia & Central Asia–North	828
Nepal & the Himalayan Foothills	394
New Zealand Maori	223
Nigeria	587
Nigeria—East Central	461
Nilotic Peoples	243
Northern Africa	438
Northern Asia	41
Northern China	245
Northern India	803
Northern Italy	1157
Northern Philippines	295
Norway	829
Portugal	1207
Samoa	112
Sardinia	117
Scotland	1737
Senegal	186
Somalia	44
Southern Bantu Peoples	220
Southern China	278
Southern India	122
Southern Italy	1651
Southern Japanese Islands	79
Southwestern China	275

Spain	965
Sweden & Denmark	1289
The Balkans	1505
Tibetan Peoples	199
Tonga	166
Vietnam	246
Wales	880
Western China	237
Western Philippines	300
Total	**71,306**

We discuss more detailed tests of the performance of the 2023 ethnicity reference panel in Section 4. For details of the method AncestryDNA uses for genetic ethnicity estimation, see Section 3.

3. AncestryDNA Ethnicity Estimation

3.1 Introduction

After establishing and validating the reference panel, the next step is to estimate a customer's ethnicity by comparing nearly 300,000 selected single nucleotide polymorphisms (SNPs) from their DNA to those of the reference panel. We assume that an individual's DNA is a mixture of DNA from some combination of the 88 populations represented in the reference panel. One example mixture is illustrated in Figure 3.1, where, because of recombination, a customer inherits stretches of DNA from his or her four grandparents who, in this example, each 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 loci in the genome were inherited from the same person and so the same population (for more details on DNA inheritance see our DNA Matching White Paper). 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. Our method takes advantage of this to greatly improve our estimates.

Our approach for estimating a customer's genetic ethnicity assumes that each segment of their genome comes from one of the 88 populations in the reference panel. We divide the customer's genome into 1,001 windows, and our approach assumes that the DNA inherited from each parent in each window comes from exactly one population (the windows are small enough that this will almost always be true). We compare the customer's DNA to that of our reference panel in each window, and combine information from all the windows to estimate what overall portion of the customer's genome came from each population using a hidden Markov model (HMM), described in Sections 3.3-3.5 below.

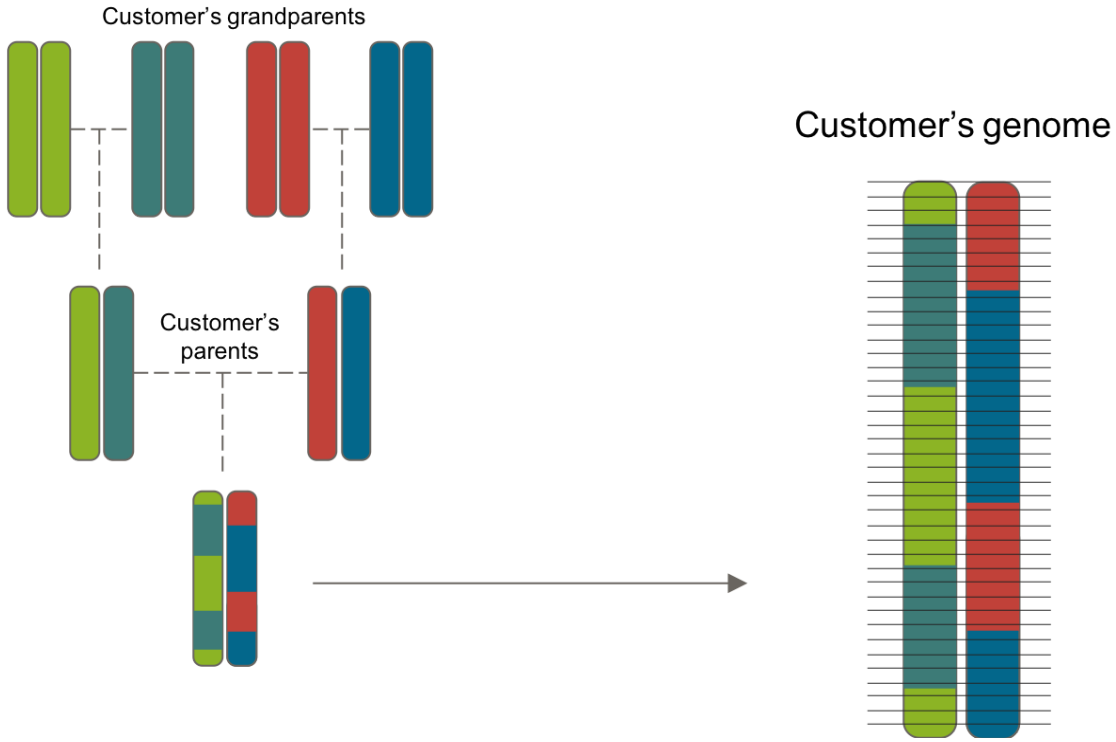


Figure 3.1: Inheritance of DNA from different populations. *On the left, we present a three-generation genetic family tree. For each individual, we show two vertical bars representing the two copies of a single chromosome present in each individual. These bars are colored to show the reference population 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.*

3.2 Phasing SNP Data

At AncestryDNA, we use microarrays to obtain DNA data from customer samples. We look at approximately 700,000 individual locations of DNA (SNPs) on chromosomes 1-22 and the X chromosome and determine the nucleotides at each position. It is important to understand that every person inherits *two* alleles, one from each parent, at each of these 700,000 sites. For example, we may see an A and a T at position 1, a G and a G at position 2, and so on. A crucial step in ethnicity estimation is to separate which letters were inherited from different parents—a process called *phasing*. We use a new technology we call SideView™ to separate DNA inherited from each parent across the entire genome. When separated, we can estimate the genetic ethnicity of DNA inherited from each parent individually and independently using the approximately 300,000 SNPs that are shared with all members of the reference panel.

SideView™ uses DNA shared with distant relatives across the genome to aid in the phasing. The correctness of the DNA phasing for an individual therefore relies, in part, on that person sharing enough DNA with other people in our database. Since this is not always the case, we design the hidden Markov model (HMM) we use for ethnicity estimation to allow for incorrect phasing, although doing so complicates the model significantly. In the next section, we explain how an HMM is useful in ethnicity estimation, first with a model to analyze one parent individually, and then we show how we extend that model to account for phase error.

3.3 Principles of a Hidden Markov Model

Our goal is to assign one of the populations from our reference panel to each window of the genome and to each parent (i.e., the DNA a customer inherited from each parent). A hidden Markov model is well-suited for this task because its strength is that it can represent thousands of interrelated variables but still perform efficient inference—using a technique called dynamic programming—as long as each variable depends on only a few others. An HMM is a set of *states* and *transitions* connected as a directed acyclic graph (the transitions move forward along the genome and never cycle back). Each transition is associated with a probability, and each state has an emission probability, which allows the HMM to compute the *posterior* probability (i.e.*,* taking all populations and windows into account) of individual states, individual transitions, and *paths* through the model. Figure 3.2 illustrates an HMM representing the DNA inherited from one parent for three reference populations (represented by green, yellow, and red)

and six windows (our complete analysis uses 88 populations and 1,001 windows). It also shows a *path* through the model (the thick blue transitions). We use HMMs to infer the most likely path (called the *Viterbi* path), which assigns exactly one population to each window of the genome. We also use HMMs to take *path samples*—alternative paths that are also likely—to get a better idea of how much the assignment to each population might vary according to the model.

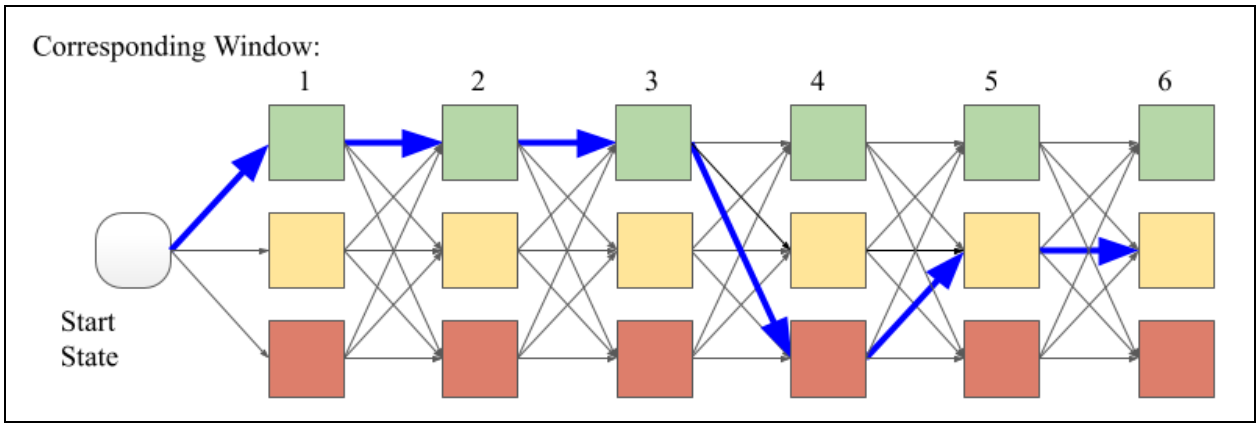


Figure 3.2: The *states* and *transitions* of an HMM representing the possible populations that explain the DNA inherited from one parent in each of several windows. *This illustration includes three populations (green, yellow, and red), and six windows. The arrows represent transitions between states, and each transition will have an associated probability. By using the transition probabilities, an HMM can compute the likelihood of each of these states and determine the most likely path through the model (illustrated by the bold blue arrows), which assigns one population to each window across the genome.*

The transition probabilities in this HMM depend on how often a population assignment should change, and, when they do change, how likely the new population is to be chosen. A transition to the same population is generally more probable in our model because the population that explains the DNA inherited from a parent is likely to be the same for several consecutive windows. However, the number of populations varies from person to person. Our HMM learns the probability of changing population states from the genotype data. When a transition does change populations, the transition probability depends also on the proportion throughout the genome of the population being transitioned to, which our approach also learns for each individual person.

The state emission probabilities in this HMM depend on the similarity between the DNA inherited from the parent and that of a reference panel corresponding to the population the state represents. We describe how we measure this similarity in Section 3.4 below.

3.4 Emission Probabilities

Determining how likely the DNA in a window came from a population (the emission probability) is a complicated process and is described in more detail in our paper Ancestry Inference Using Reference Labeled Clusters of Haplotypes.

Briefly, our approach includes the following steps:

I. **Create haplotype models for each window.** Using an ethnically-diverse set of about 50,000 individuals, we infer *BEAGLE* (Browning 2007) haplotype cluster models for each window.
II. **Annotate the reference panel.** The states in the BEAGLE models represent clusters of similar haplotypes. We wish to associate those clusters with populations. Because we are confident in the geographic origin of members of the reference panel, we are able to calculate the probability that a haplotype from a given population is represented by a particular haplotype cluster.
III. **Assign haplotype clusters to the test sample and aggregate the annotations.** Given a phased customer genotype, we observe which haplotype clusters the genotype belongs to and base the emission probabilities for a population on the weighted average annotation (how often the population reference panel belongs to the haplotype cluster, weighted so that each SNP in the window contributes equally).

HMMs are used in a number of existing approaches for estimating ancestral proportions (Maples 2013). The key part of our method is step III, where we use rich haplotype models in each window, annotated with population labels from the haplotypes in our reference panel, to assign a likelihood over all population labels to the haplotypes in our test sample. It is worth noting that our method lends itself to high-throughput ethnicity estimation, as steps I through III above—learning the haplotype models from a large training set and then annotating them with the reference panel populations—need only be carried out once.

3.5 Accounting for Phase Error

We use the HMM described above (Figure 3.2) to identify populations whose probability of assignment is virtually zero for one parent or the other, and we remove those from further consideration, but our final estimates are based on a more complicated HMM that simultaneously explains *both* haplotypes inherited from the parents. We need this more complicated model because we cannot be certain that every

genome is completely separated into DNA inherited from each parent, since SideView™ cannot phase in places where an individual has no DNA matches.

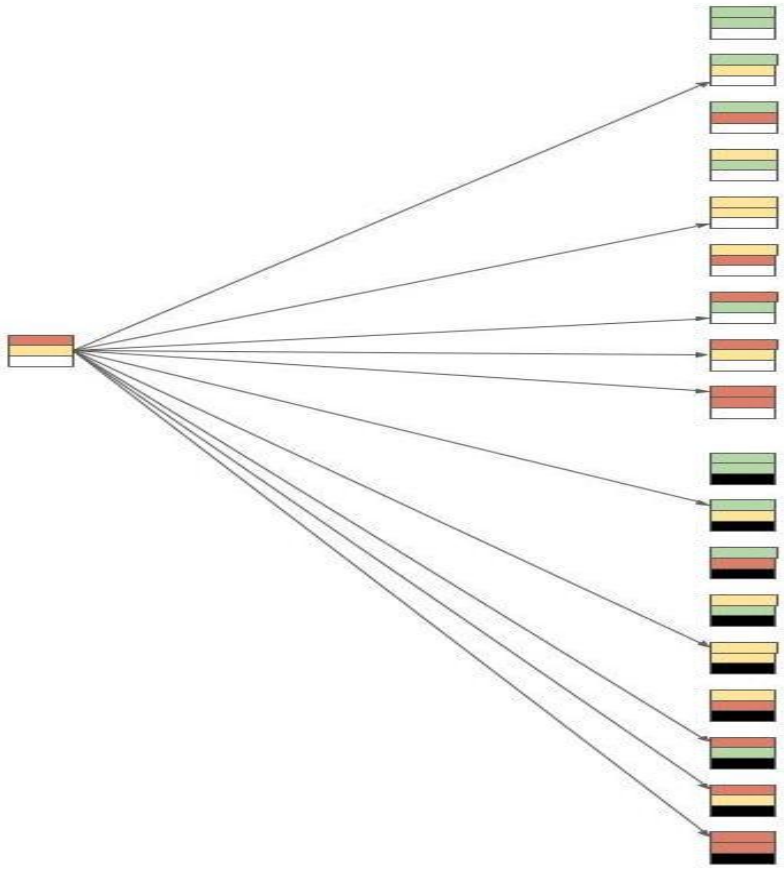


Figure 3.3: State transitions in an HMM representing *K*=3 populations. *The HMM we use in practice explains the DNA inherited from both parents simultaneously. This figure illustrates the states in a model with the same three (green, yellow, red) populations as the HMM in Figure 3.2. There are KxKx2 states in each window. Each state represents the population inherited from parent 1 (top color of each state), parent 2 (middle), and whether or not parent 1 corresponds to haplotype 1 (bottom). Only one state is shown on the left, and possible transitions to all states in the next window (right). We only consider states such that the DNA inherited from at least one parent keeps the same population assignment.*

Figure 3.3 shows the set of states necessary for the HMM we use. Each state represents the population that explains the DNA inherited from *both* parents, and we also assign one parent to haplotype 1 in the phased data and the other parent to haplotype 2 and allow those phase assignments to change from window to window. The resulting HMM has many more states, and each state represents the population that explains parent #1's DNA (*K* possible values, if there are *K* populations), the population that explains

parent #2's DNA (K possible values), and which haplotype corresponds to which parent (2 possible values). The HMM has $K \times K \times 2$ states for each genomic window and all possible transitions between them such that, at most, one parent's state changes population. While the constraint to one parent changing populations is consistent with biology—recombination events in different parents are independent—it is put in place mostly for practical reasons of efficient inference. The transition probability in this HMM (Figure 3.3) depends on two additional variables: the probability of changing phase from window to window and the probability of changing back. These values are also learned for each individual.

The parameters of the HMM are set based on several iterations of an expectation-maximization algorithm based on a standard HMM learning approach called Baum-Welch. For each individual, the algorithm learns (i) the probability of changing populations (for each parent), (ii) the overall distribution of population assignments (for each parent), (iii) the probability of changing phase (and changing back). The emission probabilities for each state are fixed throughout the process. Although the model allows for phase error, the model most often learns that the optimal estimate includes no phase corrections, and therefore the estimates for most Ancestry DNA customers are based on the SideView™ phase and parent assignments exactly.

After learning, we are able to compute through our HMM model:

1. The *Viterbi* path through the model. This is the single most likely path, according to the parameters of the model, which assigns one population to the DNA inherited from each parent in each window of the genome.
2. Probabilistic path samples through the model. These paths also assign one population to each parent in each window, and they are only slightly less likely (according to the model) than the Viterbi path, so they help describe how much or how little of a given population may still be consistent with the individual's DNA.

We report the sum population assignment for each parent according to the most likely path and report a most probable range based on 1,000 path samples taken from the model. These ranges are adjusted based on how well path samples perform on test data (see Section 4.4).

4. Assessing Ethnicity Estimation Performance

After developing and optimizing both the estimation process and the reference panel, the final step is to determine how well they perform together at assigning ethnicity. Basically, we see how close our process gets to the right answer through rigorous testing using a wide variety of test cases with known ethnicity.

4.1 Cross-Validation

We evaluate the performance of the ethnicity estimation process by running it on two different test cases where we know what the correct answer should be: single-origin individuals (including synthetic single-origin individuals) from the reference panel and simulated individuals with mixed ethnicities. We gauge its effectiveness by seeing how close we get to the true ethnicity.

Single-origin individuals: We use two different sets of *single-origin individuals* in our cross-validation studies. The first are those for whom we utilize their entire genome as a reference for a particular region. These people have a long family history in a single region and represent a typical person from that region. By definition, these individuals in our reference panel each have 100% of a single ethnicity.

This approach does not work for the reference panel regions where we used the indigenous DNA of admixed individuals. For these reference panel regions, primarily from the Americas and Oceania, we created *synthetic single-origin individuals* by piecing together genotype sequences that represent indigenous ancestry from multiple individuals. These synthetic single-origin individuals are then used to evaluate the accuracy of our method.

We evaluate our process by running 20-fold cross-validation experiments using these single-origin individuals from our reference panel. For example, if we had 100 people in each reference panel group, we would take 5 from each of the 88 groups and run the algorithm on these 440 samples using the remaining 8,360 individuals as the reference group. Then a different 5 would be taken from each group and the process repeated 20 times so that every individual in the reference panel is tested.

Overall we observe that the updated process correctly assigns an average of 85.5% of the genetic ethnicity to the correct region for single-origin individuals from our reference panel (Figure 2.3). We predicted at least 97% of the genetic ethnicity from the correct region for the following groups:

- Aboriginal & Torres Strait Islander
- Cameroon, Congo & Western Bantu Peoples
- Eastern European Roma
- Guam
- Indigenous Americas–Bolivia & Peru
- Indigenous Americas–Colombia & Venezuela
- Indigenous Americas–Mexico
- Indigenous Americas–North
- Indigenous Americas–Yucatan Peninsula
- Indigenous Americas—Chile
- Indigenous Americas—Ecuador
- Indigenous Arctic
- Indigenous Cuba
- Indigenous Haiti & Dominican Republic
- Indigenous Puerto Rico
- Jewish
- Khoisan, Aka & Mbuti Peoples
- Melanesia
- New Zealand Maori
- Southern Bantu Peoples
- Tibetan Peoples

For some regions, such as Bengal and France, the numbers are not as high, with average assignments of 56% and 58% to the correct region, respectively. However, even if the prediction accuracy falls short of 100% for some regions, the remaining ethnicity is still assigned to nearby regions. For example, individuals from Bengal might get some assignments to Southern India, and individuals from France might get some level of assignment to England & Northwestern Europe (see Figure 4.1)

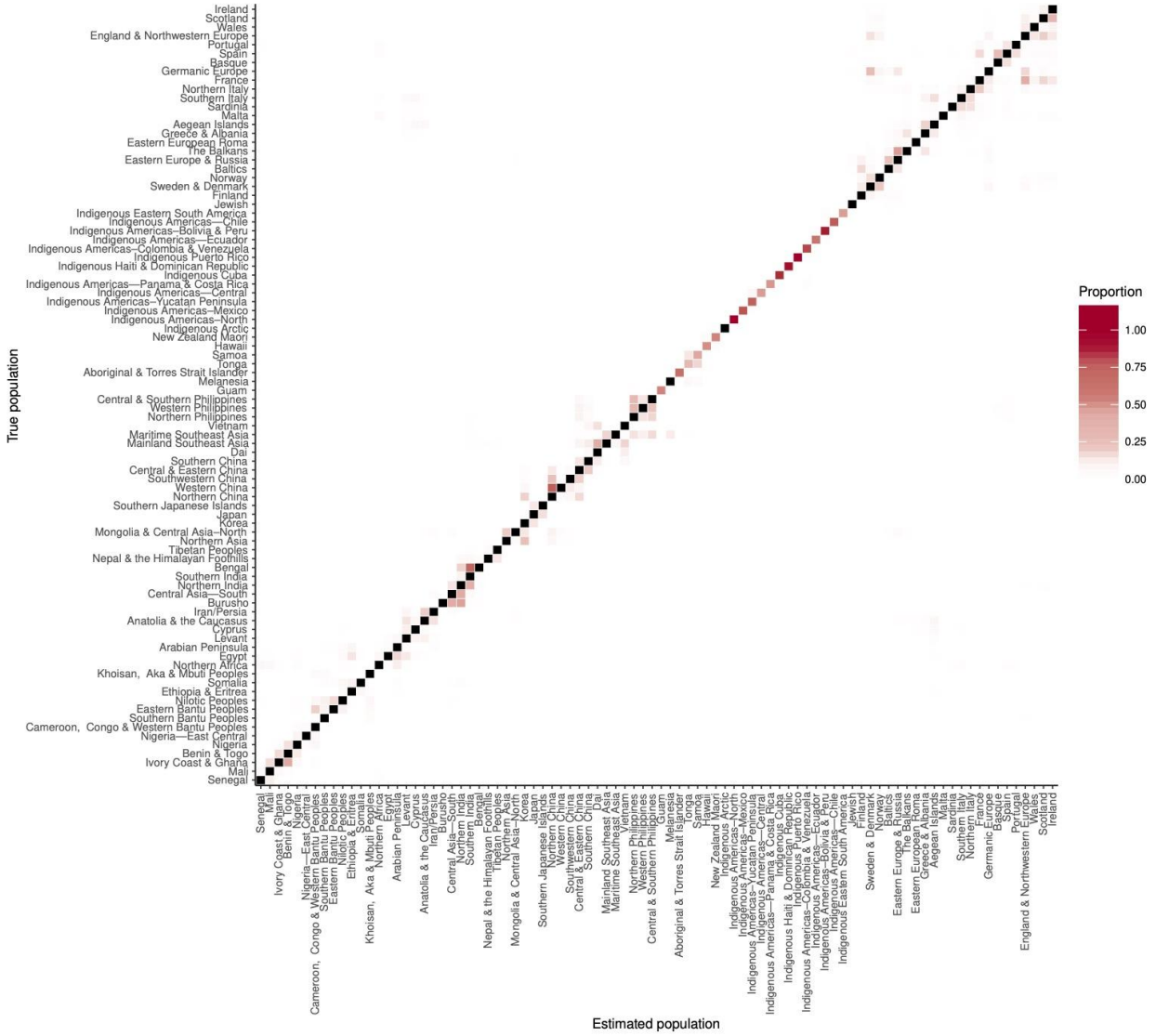


Figure 4.1: Average estimated ethnicities for single-origin individuals from each population. *In this graph, each row represents single-origin individuals from the population listed. Each column represents each of the possible 88 ethnicities that the single-origin individual might be assigned to. The graph is set up such that the matching individual and his or her ethnicity are aligned along the diagonal line. If the algorithm worked perfectly, there would be only dark red (appearing black) boxes along the diagonal—red represents 100% origin from that population. Any boxes that are not on the diagonal represent misassigned populations. This graph also shows that certain ethnicities can be confounded by other ethnicities. For example, individuals with 100% Germanic Europe ethnicity can be assigned to England & Northwestern Europe and Sweden & Denmark.*

Simulated individuals with mixed ethnicities: We also evaluated the accuracy of ethnicity estimates for "simulated" individuals of mixed ethnicity origins. These test cases are simulations we construct with known mixtures of ethnicities. Each simulated, admixed individual can have as few as 2 or as many as 32 ethnicity regions, with various proportions. Since the true ethnicity proportions are known, we can calculate precision and recall for each ethnicity region. Precision and recall are two important factors in evaluating our estimation process.

Precision can be thought of as how much of the reported ethnicity is true. For example, if our process predicts an individual has 40% Northern Africa, but only 30% really is, then the process has a precision of 0.75 for Northern Africa ethnicity. Mathematically, precision is expressed as the amount of correctly identified ethnicity divided by the estimated value for that region.

Recall can be thought of as how much of the true ethnicity is called by the process. Keeping with our Northern Africa ethnicity, imagine that an individual has 50% Northern Africa ancestry, but the algorithm predicts 40%. In this case, the process has a recall of 0.8 for Northern Africa ethnicity.

With SideView™ technology we can also calculate the precision and recall for attributing ethnicity regions to each side of your family tree (ethnicity inheritance). For example, if our process predicts that an individual inherited 100% of their total Northern Africa ancestry from one side of their family, when in reality they inherited 70% from one side and 30% from the other side, the ethnicity inheritance precision would be calculated as 0.7.

Table 4.2 : Precision/Recall for each region calculated from ethnicity estimates of simulated individuals with mixed ethnicities.

Region	**Precision**	**Recall**	**Ethnicity Inheritance Precision**	**Ethnicity Inheritance Recall**
Aboriginal & Torres Strait Islander	0.98	0.98	0.98	0.98
Aegean Islands	0.81	0.83	0.79	0.82
Anatolia & the Caucasus	0.63	0.76	0.63	0.75
Arabian Peninsula	0.78	0.93	0.77	0.93
Baltics	0.4	0.85	0.39	0.84
Basque	0.38	0.93	0.38	0.92
Bengal	0.95	0.51	0.94	0.51
Benin & Togo	0.73	0.88	0.73	0.87
Burusho	0.95	0.57	0.95	0.57
Cameroon, Congo & Western Bantu Peoples	0.91	0.97	0.91	0.97
Central & Eastern China	0.73	0.79	0.73	0.79
Central & Southern Philippines	0.76	0.79	0.76	0.78

Central Asia—South	0.88	0.76	0.87	0.76
Cyprus	0.73	0.89	0.72	0.88
Dai	0.4	0.92	0.40	0.92
Eastern Bantu Peoples	0.87	0.85	0.87	0.85
Eastern Europe & Russia	0.74	0.82	0.73	0.80
Eastern European Roma	0.93	0.97	0.93	0.96
Egypt	0.98	0.75	0.97	0.75
England & Northwestern Europe	0.52	0.72	0.48	0.67
Ethiopia & Eritrea	0.81	0.96	0.81	0.96
Finland	0.73	0.96	0.73	0.96
France	0.92	0.51	0.89	0.50
Germanic Europe	0.83	0.65	0.80	0.63
Greece & Albania	0.68	0.83	0.67	0.81
Guam	0.98	0.98	0.98	0.97
Hawaii	0.97	0.96	0.97	0.96
Indigenous Americas–Bolivia & Peru	0.94	0.99	0.94	0.99
Indigenous Americas–Colombia & Venezuela	0.97	0.98	0.97	0.98
Indigenous Americas–Mexico	0.97	0.98	0.96	0.97
Indigenous Americas–North	0.97	0.99	0.97	0.99
Indigenous Americas–Yucatan Peninsula	0.86	0.98	0.86	0.97
Indigenous Americas—Central	0.99	0.92	0.98	0.92
Indigenous Americas—Chile	0.98	0.98	0.97	0.98
Indigenous Americas—Ecuador	0.98	0.97	0.98	0.97
Indigenous Americas—Panama & Costa Rica	0.94	0.95	0.94	0.94
Indigenous Arctic	0.97	0.99	0.97	0.99
Indigenous Cuba	0.99	0.99	0.98	0.99
Indigenous Eastern South America	0.99	0.94	0.98	0.95
Indigenous Haiti & Dominican Republic	0.98	0.99	0.98	0.99
Indigenous Puerto Rico	0.98	0.99	0.98	1.00
Iran/Persia	0.96	0.78	0.95	0.78
Ireland	0.48	0.95	0.47	0.93
Ivory Coast & Ghana	0.84	0.72	0.84	0.72

Japan	0.87	0.93	0.87	0.92
Jewish	0.92	0.99	0.92	0.98
Khoisan, Aka & Mbuti Peoples	0.87	0.98	0.87	0.98
Korea	0.75	0.93	0.75	0.92
Levant	0.63	0.88	0.63	0.87
Mainland Southeast Asia	0.96	0.63	0.96	0.63
Mali	0.91	0.96	0.91	0.96
Malta	0.91	0.88	0.90	0.88
Maritime Southeast Asia	0.87	0.66	0.87	0.66
Melanesia	0.92	0.98	0.92	0.98
Mongolia & Central Asia–North	0.99	0.77	0.98	0.77
Nepal & the Himalayan Foothills	0.98	0.9	0.98	0.90
New Zealand Maori	0.93	0.97	0.93	0.97
Nigeria	0.93	0.9	0.93	0.89
Nigeria—East Central	0.95	0.96	0.95	0.96
Nilotic Peoples	0.93	0.85	0.93	0.84
Northern Africa	0.91	0.9	0.91	0.90
Northern Asia	0.3	0.81	0.30	0.81
Northern China	0.53	0.75	0.52	0.74
Northern India	0.78	0.75	0.77	0.74
Northern Italy	0.78	0.66	0.76	0.64
Northern Philippines	0.74	0.79	0.73	0.79
Norway	0.64	0.9	0.62	0.88
Portugal	0.92	0.8	0.91	0.79
Samoa	0.65	0.8	0.65	0.80
Sardinia	0.63	0.81	0.62	0.80
Scotland	0.59	0.81	0.57	0.77
Senegal	0.88	0.96	0.88	0.96
Somalia	0.89	0.93	0.89	0.93
Southern Bantu Peoples	0.92	0.98	0.92	0.98
Southern China	0.74	0.86	0.74	0.86
Southern India	0.31	0.95	0.31	0.94
Southern Italy	0.94	0.65	0.93	0.65

Southern Japanese Islands	0.83	0.96	0.83	0.95
Southwestern China	0.87	0.73	0.87	0.73
Spain	0.83	0.61	0.81	0.61
Sweden & Denmark	0.55	0.83	0.53	0.80
The Balkans	0.92	0.65	0.90	0.64
Tibetan Peoples	0.89	0.98	0.89	0.97
Tonga	0.8	0.69	0.80	0.69
Vietnam	0.78	0.89	0.77	0.89
Wales	0.65	0.93	0.64	0.91
Western China	0.92	0.54	0.91	0.54
Western Philippines	0.83	0.7	0.83	0.70

We found that the majority of our ethnicity regions have precision and recall that are both greater than 0.75. The following regions have both precision and recall exceeding 0.9:

- Aboriginal & Torres Strait Islander
- Cameroon, Congo & Western Bantu Peoples
- Eastern European Roma
- Guam
- Hawaii
- Indigenous Americas–Bolivia & Peru
- Indigenous Americas–Colombia & Venezuela
- Indigenous Americas–Mexico
- Indigenous Americas–North
- Indigenous Americas—Central
- Indigenous Americas—Chile
- Indigenous Americas—Ecuador
- Indigenous Americas—Panama & Costa Rica
- Indigenous Arctic
- Indigenous Cuba
- Indigenous Eastern South America
- Indigenous Haiti & Dominican Republic
- Indigenous Puerto Rico

- Jewish
- Mali
- Melanesia
- New Zealand Maori
- Nigeria—East Central
- Southern Bantu Peoples

Some regions, such as France and Bengal, have relatively lower recall, both 0.51. Some regions have relatively low precision, such as Southern India (0.31), Northern Asia (0.30), Basque (0.38), Dai (0.40), and Baltics (0.40).

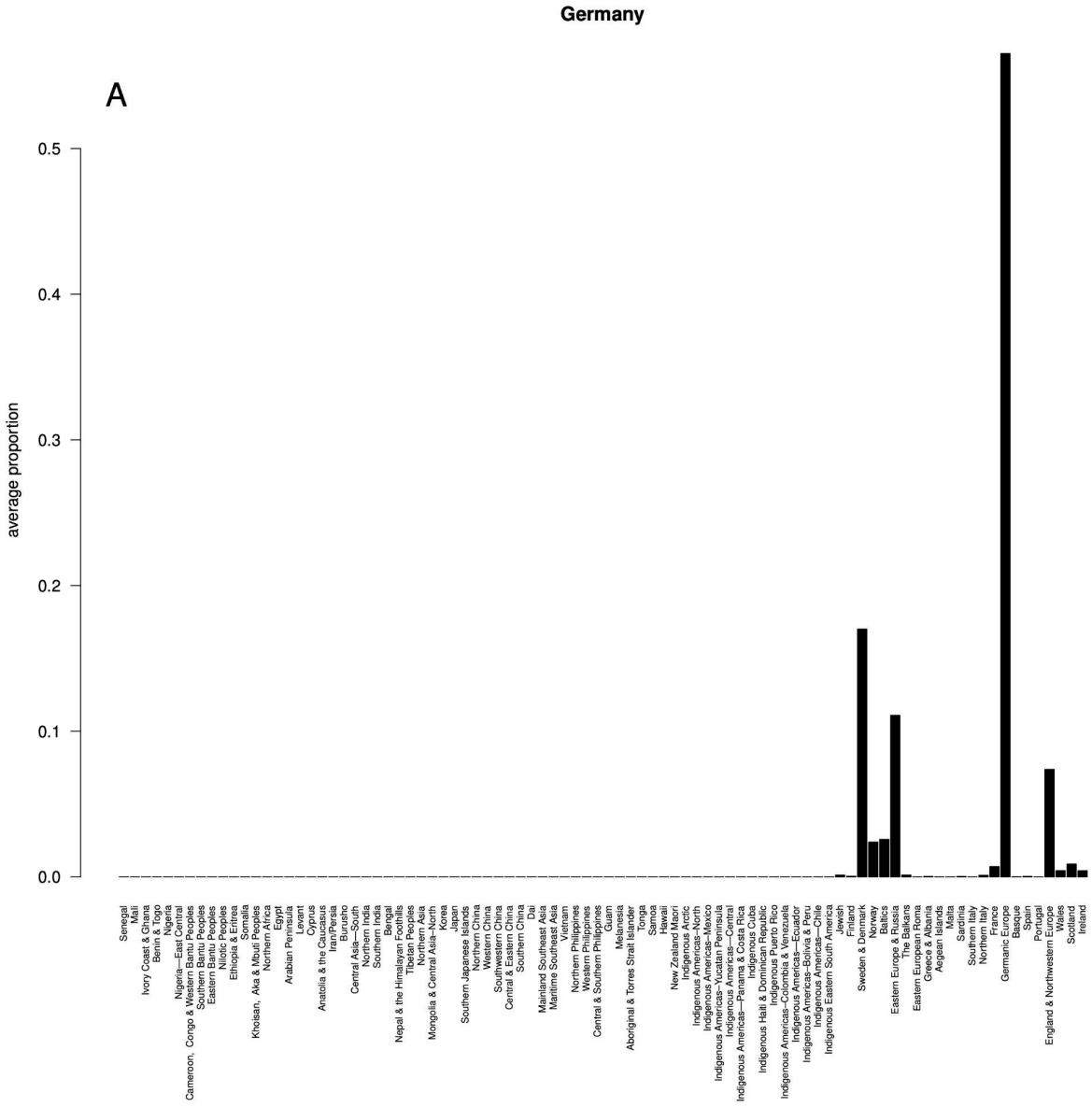
Ethnicity inheritance precision and recall are slightly lower on average when compared with precision and recall when both sides of the genome are considered together, 81.4 vs 81.9 and 84.9 vs 85.5 respectively. This is because there will always be some additional error when trying to estimate which side of your family tree you inherited a particular ethnicity region from. In general, parsing your inheritance patterns by one side of your family tree versus the other is a more difficult problem than when both sides of the tree are considered together.

4.2 Region Assessment

To assess how our regions perform in different parts of the world and to help our customers interpret their results, we can look at the ethnicity estimates of people with deep genealogical roots back to the same country or part of a country who are not included in our reference panel. To find these individuals, we use customer-created family trees and look for customers who have consented to research and have all of their ancestors from the same country. Ideally, we'd only look at people with all of their grandparents (or older) from the same country, but due to low numbers for some countries we sometimes include people where only their parents are from the same country.

Customers who are not in the reference panel and have deep trees tracing back to a single country are expected to have high assignments to the regions associated with that country, and this is what we generally find. For example, Figure 4.3A shows the average ethnicity assignments for 200 customers with all four grandparents (or older) born in Germany. As you can see, while most of their assignment is to the Germanic Europe region, other regions do appear in small but significant amounts. Figure 4.3B shows a

similar pattern for 200 customers with all four grandparents born in Korea. These analyses help ensure that ethnicity estimates for people from a region agree with expectations.



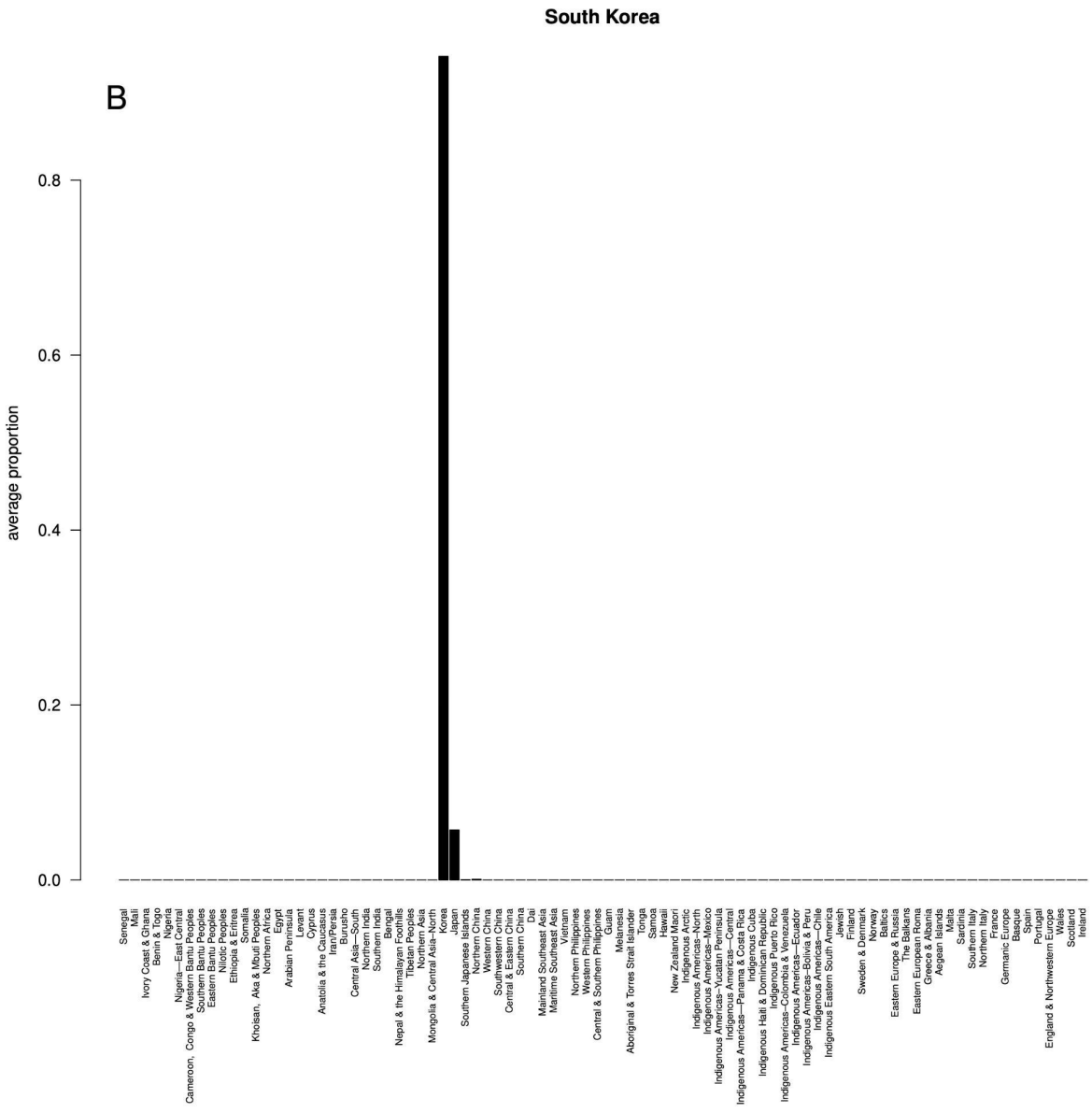


Figure 4.3 Average ethnicity assignments based on grandparents' birth location. *Average ethnicity assignments for customers with all four grandparents born in the same country. (A) Germany, (B) South Korea*

We also use maps like the one shown in Figure 4.4 to ensure that ethnicity estimates make sense geographically. The geographic distribution and amount of ethnicity estimates within a country can often help make sense of otherwise surprising results. For example, in Figure 4.4 there is assignment to the Scotland region in the Brittany area of France. This makes sense because of shared Celtic ancestry in

both Scotland and Brittany, where the Celtic language Breton is traditionally spoken. Scotland estimates in England also likely reflect the history of Celtic ancestry in that area.

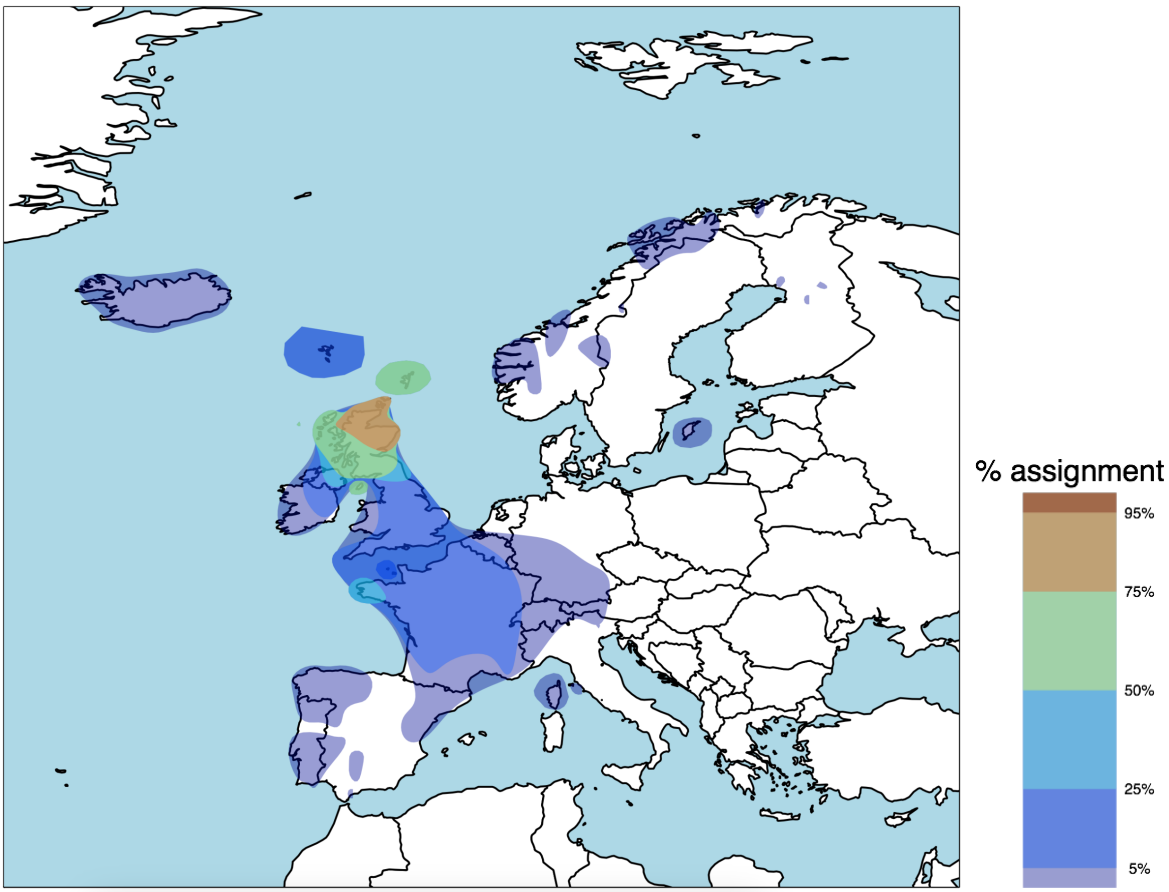
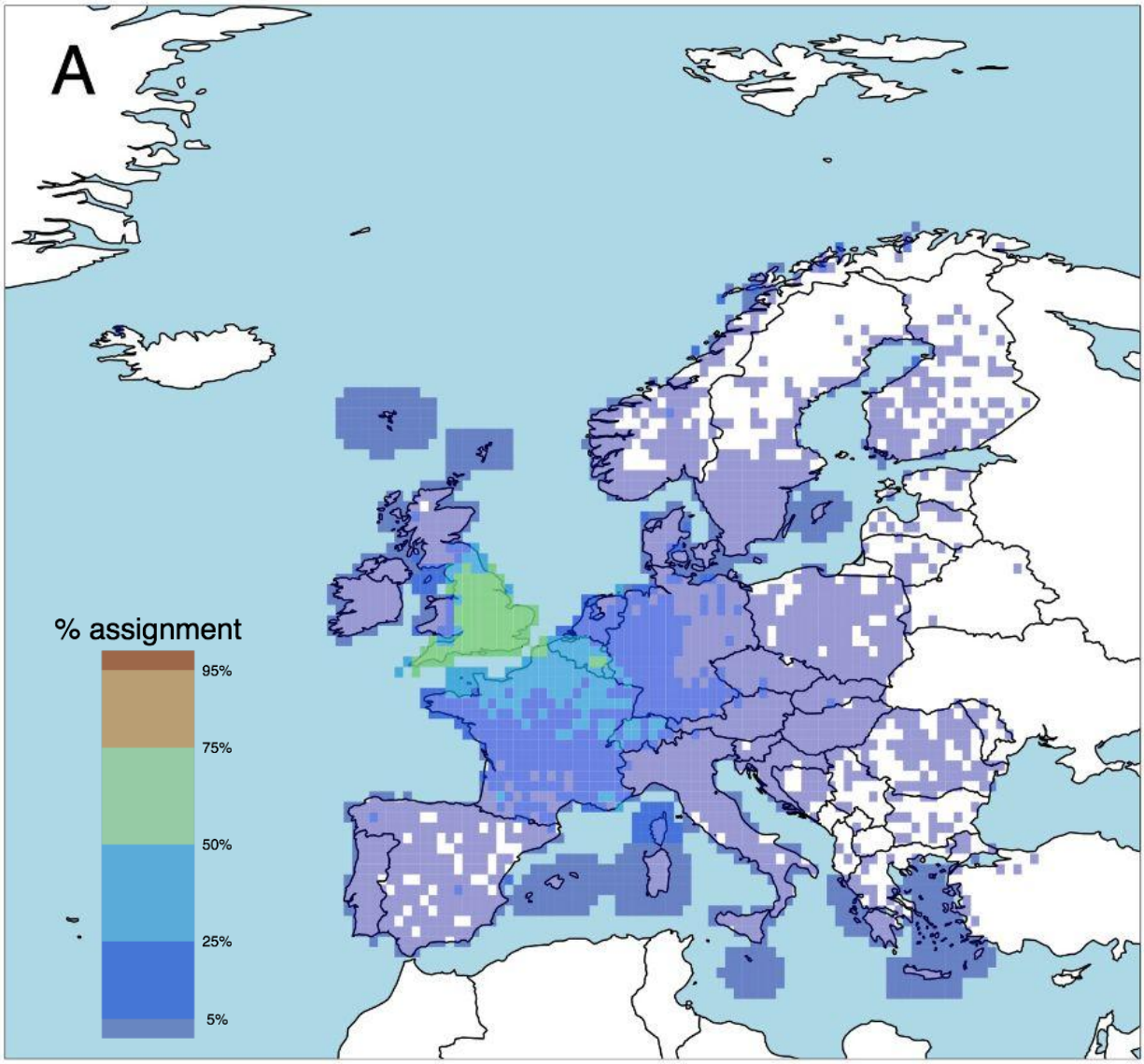


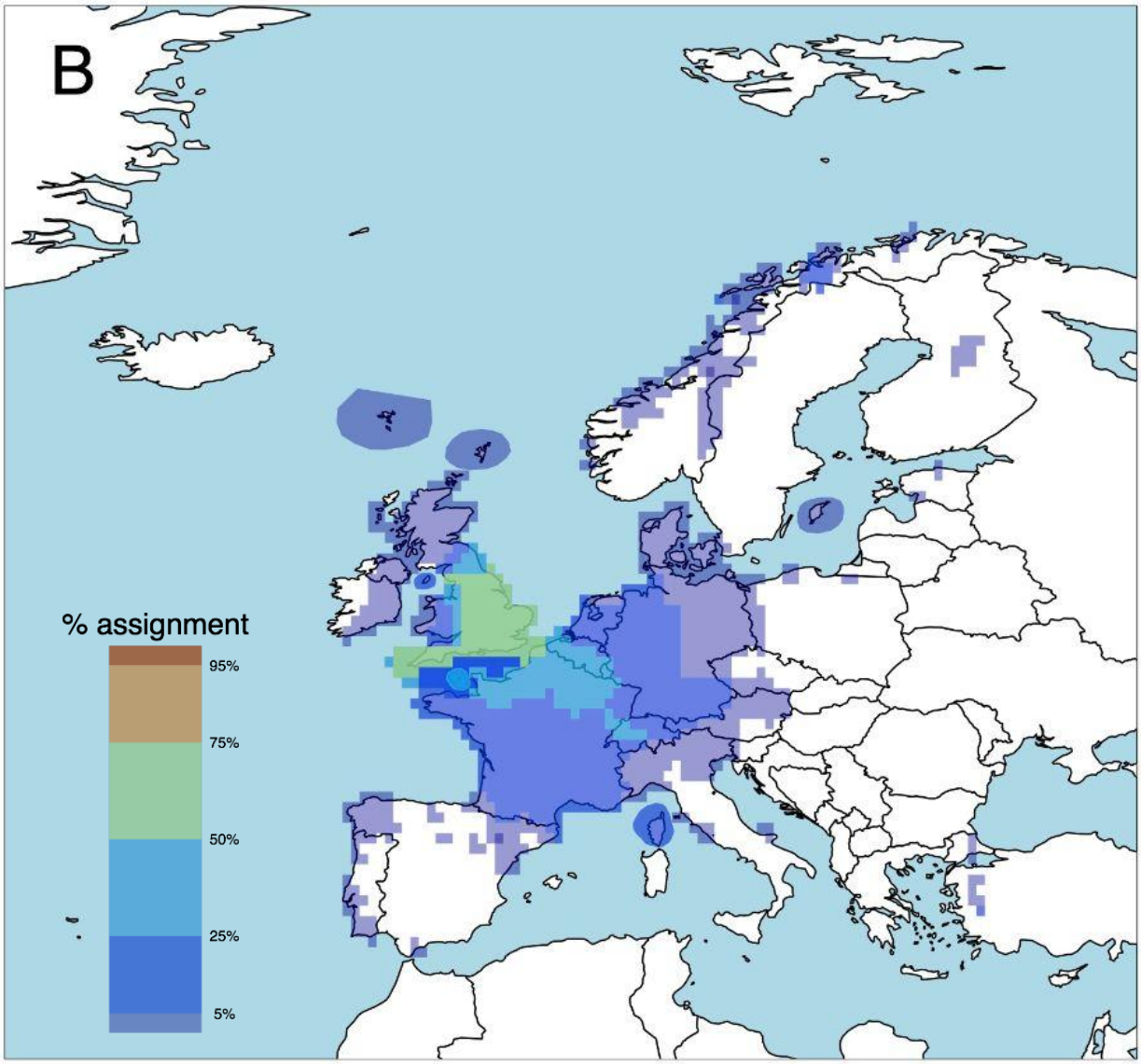
Figure 4.4 Map of average Scotland estimates. *Average estimates between 5% and 25% (darker blue) in Wales, England, and Ireland and between 25% and 50% (light blue) in Brittany likely reflect shared Celtic ancestry.*

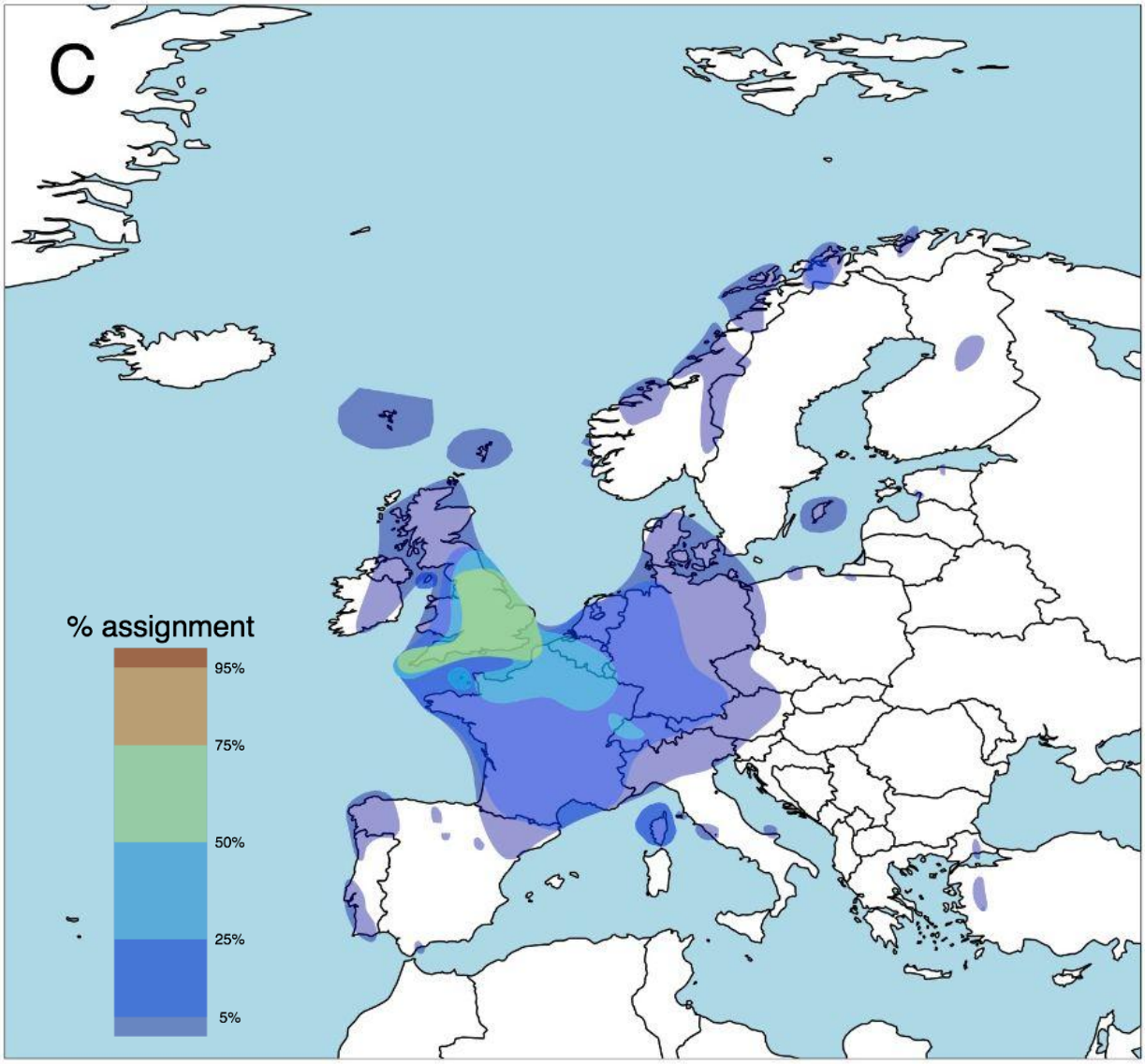
These analyses help us understand the genetic diversity of the regions and allow us to better communicate these results to our customers (e.g., even if all of a customer's ancestors are German, the customer can expect to see additional regions in their ethnicity estimate). These analyses also aid us in prioritizing future developments for further ethnicity estimation updates.

4.3 Regional Polygon Construction

We divide the world into 88 regions in our reference panel. Each region represents a population or group of related populations with a unique genetic profile reflecting their shared ancestry. Where possible, we use the known geographic locations of our samples to guide how we create the regions. Figure 4.5 shows an example of the ethnicity estimate and geographic information used to define the polygon for our England & Northwestern Europe region.




B
% assignment
95%
75%
50%
25%
5%


C
% assignment
95%
75%
50%
25%
5%

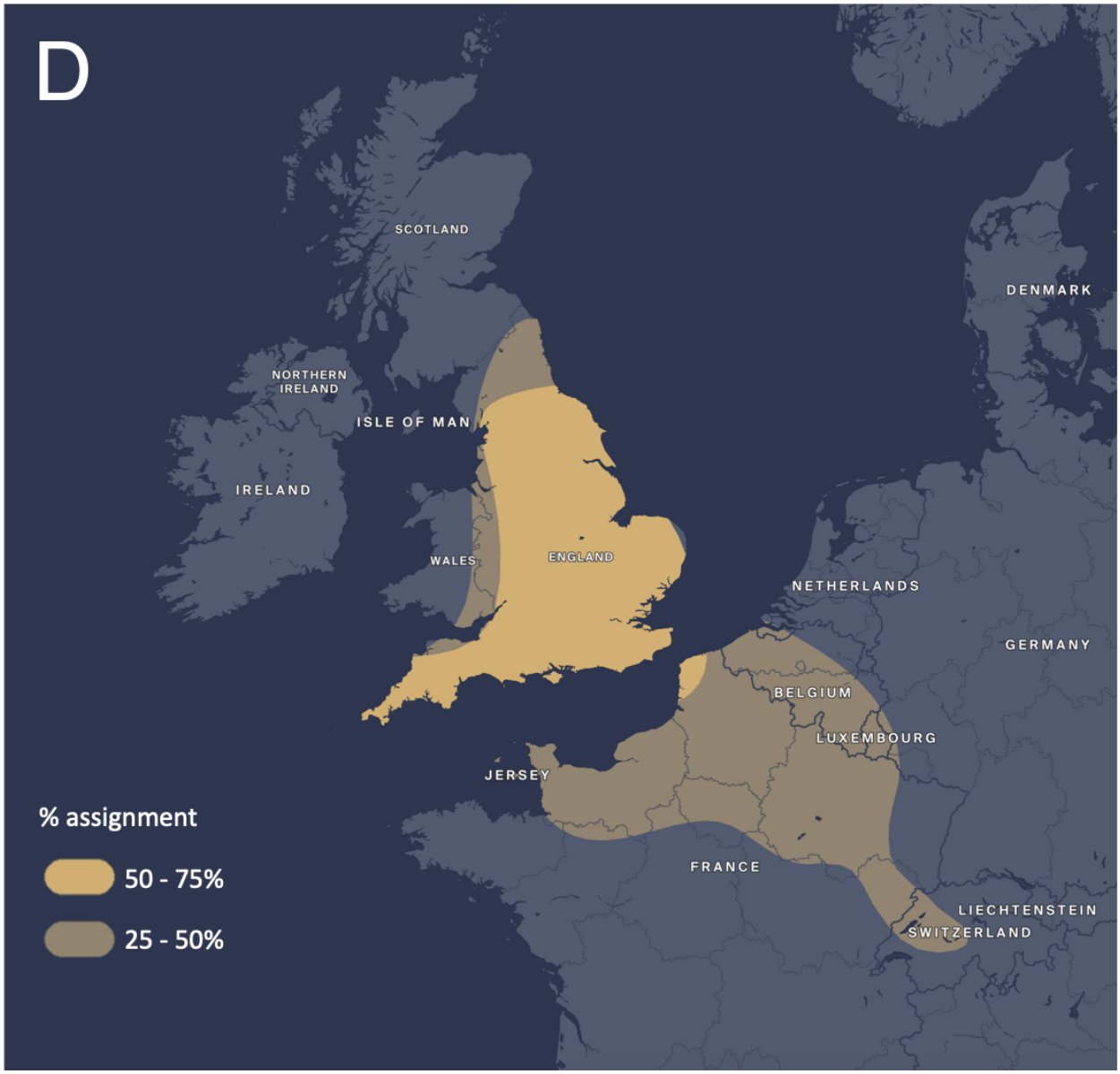


Figure 4.5: Using geographical sample locations to draw regional polygons. *Panel A shows the distribution of the England & Northwestern Europe ethnicity predicted for a set of samples with geographic information. Samples are assigned to grids of 0.5 degrees longitude by 0.5 degrees latitude based on the average birth location of their ancestors' grandparents. The color of each grid squarepoint on the map represents the average England & Northwestern Europe ethnicity of samples from each grid. Panel B shows the maps after filling in missing regions and smoothing the results. Panel C shows the information processed with smoothing, creating the outlines representing the ancestry regions shown to customers. Panel D shows the final product version after manual touch-ups.*

In Figure 4.5A, we show the amount of our England & Northwestern Europe ethnicity region assigned to a combination of reference panel samples, customers with deep roots from the same country, and

noncustomer samples. Figure 4.5B shows the results after imputing values to fill in gaps and applying smoothing methods to make the plot less spotty. It is clear from the plot that there is a gradient of ethnicity in this area that is centered in England that quickly tapers off in surrounding areas. For example, the next level of concentration, represented by light blue in the image, is in northeastern France, Belgium, the Netherlands, and Switzerland. The ethnicity gradient continues to diminish as represented in purple, with the borders reaching as far away as Northern Italy, Norway, and Portugal.

Manual edits are sometimes performed on polygons to better align them with geography like narrow peninsulas or when the polygons may imply finer-scale population structure than the underlying genetic data support. For example, the >25% polygon for the England & Northwestern Europe region on our website (Figure 4.5D) connects the separate polygons in France and Switzerland seen in Figure 4.5C. Additionally, polygons representing two of our ethnicity regions—Jewish and Khoisan, Aka & Mbuti Peoples—have hand-drawn components to describe minority populations that may not be explicitly defined by geography.

Importantly, these maps do not represent where we think a customer's ancestors may be from. Instead, the polygons broadly show how much assignment to each region is typically seen among people with deep roots from a given location. Polygons can be thought of as an extension of the region name but can more easily provide information than words alone. The polygons appear as nested zones with increasing depth of shading representing differences in the average level of ethnicity assignment. Each set of polygons is accompanied by a summary of the history of the region.

The map below shows polygons for all 88 ethnic groups mostly based on the polygons with 50% or more assignment, constructed as described above.

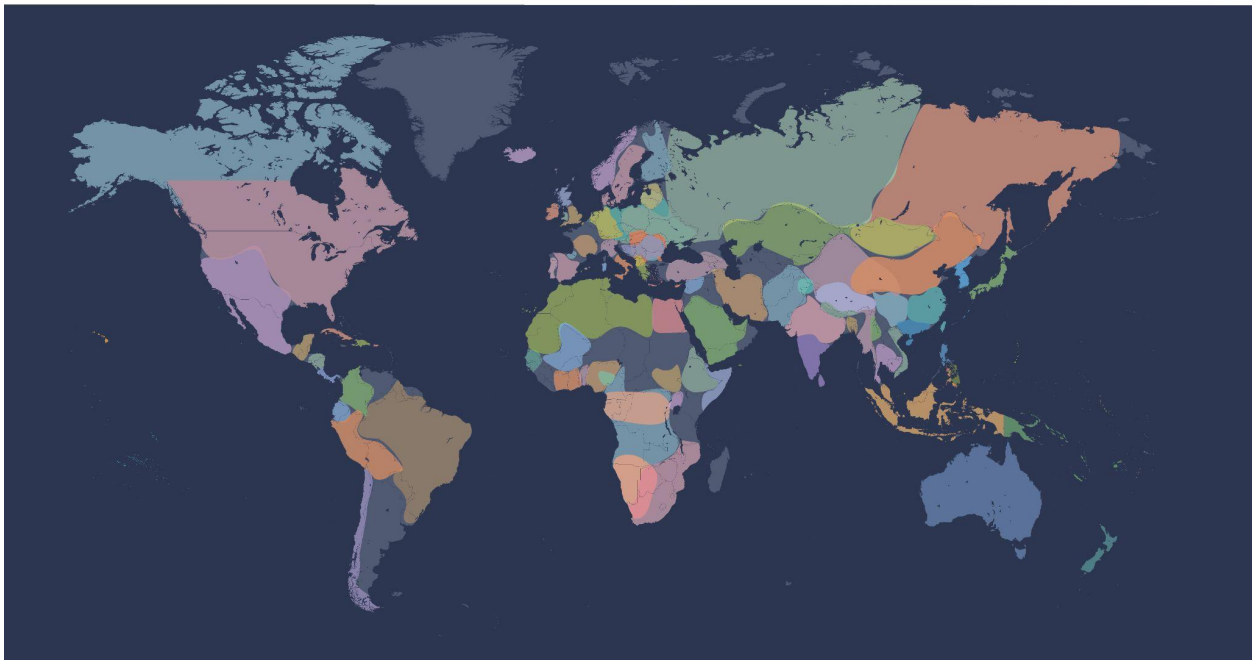


Figure 4.6: Map of 88 ethnicity region polygons.

4.4 Reporting uncertainty of estimated values

Ethnicity estimates are not an exact science. The percentage AncestryDNA reports to a customer is the most likely percentage within a range of percentages. In this section, we discuss how we calculate this range. It is important to keep in mind that at AncestryDNA we continue to build upon our previous work to offer ever more accurate results to our customers.

For example, we might report someone as 40% England & Northwestern Europe with a range of 30-60%. This means that they are most likely 40% England & Northwestern Europe but they could be anywhere between 30% and 60% England & Northwestern Europe.

As discussed in Section 3, we run a genome-wide Viterbi estimate on a customer's DNA sample and report that back as the customer's most likely ethnicity estimate. In addition to the Viterbi estimate, we are able to sample other ethnicity estimates that, while themselves likely, are not guaranteed to be the most likely estimates. The ethnicity range estimates we report are based on 1,000 samples of these non-Viterbi ethnicity estimates and correspond to the degree of uncertainty around the true ethnicity estimate. For

example, if a window has an 80% chance of being from England & Northwestern Europe, then it has a 20% chance of being from some other region. The reported range captures the uncertainty in the ethnicity estimate across a customer's DNA.

We devised a way, using the mean and standard deviation of 1,000 sampled ethnicity estimates, to estimate the range surrounding the Viterbi estimate reported to the customer. These ranges are specific to the ethnicity region in question and differ from person to person depending on their specific Viterbi estimates. Our 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.

We can test our process for calculating the range using simulated admixed individuals, like those used for the cross-validation studies, to determine how often it correctly gets the known ethnicity percentage within the range. In other words, how often does the range overlap the known ethnicity. We find that the algorithm performs very well for some populations and less well for others. Since we know the true ethnicity in the simulated admixed individuals, we can incorporate correction factors specific for each population to maximize the probability that the true ethnicity falls within the range.

A large range often reflects the challenges we face because geographically neighboring regions have similar DNA. What this means is that if a customer's ethnicity estimate includes many neighboring regions, their ranges will most likely be larger than if it contained more distant regions. For example, while we may be fairly certain that a customer has 50% Korea and 50% Portugal ancestry (and therefore small ranges), we may be less sure about a customer who gets 50% Spain and 50% Portugal. It is easier to tell Korea from Portugal, but harder to tell Portugal from Spain. This may be reflected in the larger ranges for the second customer. But it is important to keep in mind that we are very confident of the European heritage of customer two, we are just less certain about how much ancestry is derived from Portugal and how much from Spain. It is worth noting that, in general, as we increase the precision of our regions (e.g., breaking Ireland & Scotland into two separate regions), the ranges may become larger, and this is because DNA from neighboring regions is still very similar.

5. Future Ethnicity Estimation Refinement

While AncestryDNA is extremely proud of the updates in this release of its genetic ethnicity estimation process, we will continue to improve the product over time. The availability of new data, the development of new methodologies, and the discovery of new information relating to patterns of human genetic variation will all enable future improvements to the product.

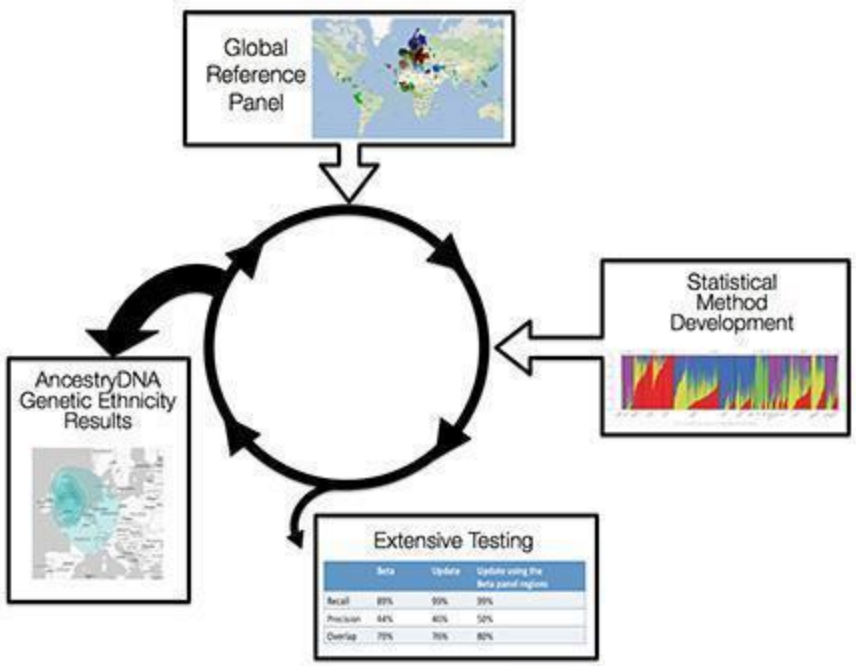


Figure 5.1: Ethnicity Improvement Cycle.

Each of the steps above represents a critical part of our ethnicity estimation procedure and development. Currently, we are working to further expand our global reference panel for future ethnicity updates. We have already begun genotyping and analyzing samples for a future update, which will provide finer-grained estimates of ethnicity. We are also continuing our diversity initiative, which gathers DNA samples from underrepresented populations around the world to expand the number of regions we can report back to customers.

Simultaneously, we are also working to improve our algorithms for ethnicity estimation. Future ethnicity updates may include improvements to our statistical methodology that will more fully leverage information in genetic data to reveal even more details about population history. Along the way, we always perform

thorough testing, involving analyses like those described above. These tests inform the focus of our improvements and help refine our improvements as necessary.

Each new release of genetic ethnicity estimation will represent a step forward in our ability to give our customers a complete description of their genetic ancestry and inform them about their genetic origins. We hope that, like the entire team at AncestryDNA, our customers will look forward to these future developments.

6. References

- AncestryDNA DNA matching white paper. https://www.ancestry.com/corporate/sites/default/files/AncestryDNA-Matching-White-Paper.pdf
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL. A human genome diversity cell line panel. Science, 2002 Apr 12;296(5566):261-2.
- Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. Nat Rev Genet. 2005 Apr;6(4):333-40.
- D.H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 2009. 19:1655–1664.
- Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genetics. 2019. 15(11): e1008432.
- International HapMap Consortium. A haplotype map of the human genome. Nature. 2005 Oct 437(7063): 1299–1320.
- International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007 Oct 449(7164):851–61.
- Jackson, J.E. A User's Guide to Principal Components (John Wiley & Sons, New York, 2003).
- K. Noto, Y. Wang, M. Barber, J. Granka, J. Byrnes, R. Curtis, N. Myres, C. Ball, and K. Chahine. Underdog: A fully-supervised phasing algorithm that learns from hundreds of thousands of samples and phases in minutes., 2014. Invited Talk at the American Society of Human Genetics (ASHG) annual meeting, San Diego, CA, October 2014.
- K. Noto, Y. Wang, M Barber, J. Byrnes, P. Carbonetto, R. Curtis, J. Granka, E. Han, A. Kermany, N. Myres, C. Ball, and K. Chahine. *Polly*: A novel approach for estimating local and global admixture proportion based on rich haplotype models. 2015. Invited Talk at the American Society of Human Genetics (ASHG) annual meeting, Baltimore, MD, October 2015.

- Lazaridis, I et. al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature, 513:409–413, 2014.
- Maples, Brian K., et al. "RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference." *The American Journal of Human Genetics* 93.2 (2013): 278-288.
- Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet 2006 2(12): e190.
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun;155(2):945-59.
- Purcell, S. PLINK v1.07. http://pngu.mgh.harvard.edu/purcell/plink/
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75.
- Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition.*Proceedings of the IEEE*, 77(2):257–286, 1989.
- S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81:1084–1096, 2007.