Ancestral Regions 2024 White Paper

Last updated October 10, 2024

Rekha Angara, Jeffrey Adrion, Ross Curtis, Kevin Keys, Jenna Lang, Keith Noto, Richard Olpin, Alisa Sedghifar, Natalie Swinford, Yong Wang, Aaron Wolf, Ju Zhang (in alphabetical order)

Summary:

Reporting ancestral regions ("regions") is one of several tools that AncestryDNA offers customers on their journey to discover their heritage, ancestors, and family history. Ancestral regions is a feature that connects users directly to the populations from which their ancestors likely came. This information can be used in conjunction with ancestral journeys discovered through **Genetic Communities**, and with relationships discovered through **DNA Matching** to better understand one's more recent and distant past.

AncestryDNA employs 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 ancestral regions. The AncestryDNA science team has advanced the science and technology behind the region results this year, producing an increase in both the overall accuracy of the results, as well as the number of regions available for assignment (from 88 regions to 107). By adding these new regions, we provide even greater granularity to our members.

This white paper will delve into the science behind:

1. How our reference panel samples are chosen and the makeup of our 107 reference panels for 2024
2. How our algorithm works to estimate a customer's genetic origins
3. Some of the results from our most recent advances for inferring ancestral origins from DNA

Glossary

Admixed — Describing an individual or population that has 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.
IBD — A term abbreviated from “Identity-by-descent.” When two individuals share DNA, we can say they have DNA that is IBD, if there is evidence that they share that DNA because they inherited it from a recent ancestor.
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. For example, we can refer to the results for a customer from our microarray as a “customer’s genotype.”
Haplotype — A stretch of DNA along a chromosome containing a group of nucleotide polymorphisms.
Hidden Markov model (HMM) — A statistical model for determining a series of hidden states based on a set of observations.
Locus/Loci — A location or locations 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.
Reference Panel — A set of people whose DNA is typical of DNA from a certain place—people native to a place or group. The DNA of these people is used as a representation of the typical DNA from this place for the purposes of studying population genetics and history.
Single nucleotide polymorphism (SNP) — A single position (nucleotide) in the genome where different variants (alleles) are seen in different people.

2. Constructing Population Reference Panels

2.1 Reference Panels are Critical to Calculate Ancestral Regions

The basic premise behind ancestral regions inference can be summarized as follows.Two **haplotypes** from the same geographic region or the same population will share more DNA with one another than will two haplotypes from different regions or groups. So two people with a historical connection to Portugal will have more DNA in common than will a person from Korea with a person from Portugal.

In practice, region inference involves comparing a person's DNA to the DNA of multiple **reference panels**. A reference panel is composed of individuals whose DNA is representative of a population. Our algorithm compares a person's DNA segments to these reference panels to determine the best-matching populations. If, for example, a section of a person's DNA looks most similar to DNA of people in our Norway reference panel, that section is said to be from Norway, and so on. The end result is a genome-wide report where individual sections of DNA are associated with one of the 107 regions in our reference panel. The similarity breakdown is provided as a total percentage breakdown, and also as a per-parent and per-chromosome report.

The accuracy of our results depends on the quality of our reference panel. Because of this, AncestryDNA has invested a significant amount of effort in collecting DNA data from populations across the globe and developing the best possible set of reference samples.

The rest of Section 2 describes the steps taken to develop our current reference panel.

2.2 Developing Reference Panels in Regions of the World with Significant Amounts of Data

AncestryDNA has genome-wide genotype data for over 25 million customers from around the world. Additionally, many people in our database have connected family trees to their DNA results, providing invaluable contextual information about the origins of these individuals. The wealth of DNA and genealogical information allows us to create robust reference panels for many global populations, especially in regions of the world where our customers' origins are concentrated.

However, it is problematic to rely solely on self-reported genealogical information from customer trees when deciding what individuals to include in the reference panel and which populations they should represent. Relying just on samples that have connected family trees would limit our number of reference panel candidates too dramatically. As well, family trees can be difficult to verify, and can carry errors.

Therefore, we've adopted a strategy that is primarily driven by genetic relatedness (IBD-sharing) among people and populations, while still incorporating information in the aggregate from ancestral region results and family trees. Specifically, we leverage our **Genetic Communities** technology and the 2,900+ DNA communities we have discovered as the bases for reference panels.

Genetic Communities is a method to identify networks of Ancestry members that are highly interconnected due to sharing DNA from a common set of ancestors. We are also able to look at family

tree information for these networks in aggregate to identify shared places of origin, patterns of movement, statistically enriched surnames, and averaged region results. These data provide insight into the specific identity and story of these genetically related groups.

Using the networks we discover as a basis for ancestral region reference panels has several benefits. First, the networks are entirely driven by genetics and self-organized population structure. Relying on patterns of genetic similarity between individuals in a group is preferable to hand-curating a set of individuals to represent a population based on self-reported data such as origins, language, or ethnicity. Second, when we do need to rely on information about origins, language, and ethnicity for identifying and annotating the reference populations, we are using information that has been averaged from hundreds or thousands of network members. By averaging the data, we remove the disproportionate effects of outliers or the need to extensively verify several hundred individual records.

A complication factor is that individuals' membership to these networks are non-exclusive. Specifically, an individual may belong to multiple different networks, representing various branches of their family tree. For example, if a person has one parent of Irish descent and another parent of Italian descent, they will likely belong to both Irish and Italian networks. Using this individual in any Italian or Irish reference panel would adversely impact the performance of our analysis.

In order to optimize our selection of reference panel candidates, we therefore adopted several filtering approaches based exclusively on genetic data.

1. We mapped the networks to corresponding world regions, and regions with rich data were selected for this reference panel development approach.
2. We considered reference panel candidates for each region and selected samples that were more likely to descend from a single origin population, i.e., do not have recently admixed family origins. We did this by filtering out individuals who had a weak genetic connection to their assigned networks (based on their number of matches to other members of that network) and who were assigned to multiple networks from different populations (e.g., Irish and Italian).
3. We filtered individuals based on their current genetic results, removing individuals with exceptionally high levels of additional off-target regions. Specifically, we identified the most common regions shared among individuals of each network, and through an iterative analysis, determined specific region and percentage thresholds to ensure robust reference panels. For example, consider that in England, most individuals with deep family roots to the south and east, and who could be high quality reference panel candidates, will likely carry between 3-5% Scandinavian ancestry. This is a result of the historical invasions of England by Germanic and

Viking tribes 1500 years ago. Despite these individuals having a genetic connection outside of England, we would still want to consider them as reference panel candidates.

4. We filtered individuals who had a low number of matches with others in the network, and who had a proportionally high number of matches to individuals in other networks. Specifically, for each sample we calculated the number of matches to others in the network and to other networks, and identified a percentile cutoff for these metrics, below which all individuals were excluded. The effect was to remove individuals who showed a low level of genetic connectedness to the region specifically and a high level of genetic connectedness to outside regions.
5. As a final check on the individuals selected for a region's reference panel, we aggregated the birth location data for these individuals and their ancestors and plotted the locations on a map. We found strong signals of enrichment for ancestor birth locations in the geographic regions of interest, and very little signal outside the region of interest.

Using the IBD-sharing and filtering approach outlined above, we identify the individuals who are best suited to include in the reference panel. Given the large number of samples available at AncestryDNA, we subsample to a maximum of 2000 individuals per reference panel to train our model ("training set"), along with 500 testing samples to tune parameters ("testing set"), and 500 samples for final validation ("validation set").

2.3 Developing Reference Panels in regions with limited data

In some regions of the world, AncestryDNA has not yet acquired enough customer samples to rely on the process described above (Section 2.2). In these regions, AncestryDNA relies on a collection of labeled samples from worldwide populations to construct training, testing, and validation sets.

We build up this collection using several different data sources including:

- 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)
- 2,500 samples from 19 populations from the 1000 Genomes Project (McVean *et al.*, 2012)
- 900 samples from 84 populations from the Human Origins dataset (Lazaridis *et al.*, Nature 2014).
- Proprietary AncestryDNA reference collections
- AncestryDNA samples from customers who consented to participate in the research project

Once this collection has been compiled, we again use unsupervised techniques that leverage shared IBD to identify groups of individuals with shared ancestry. We then use aggregated meta-data (e.g., self-reported ethnicity, spoken language) to identify the groups. We sample a subset of the total individuals to include in our reference panel.

2.3 Developing Reference Panels in Regions with significant admixture

In some parts of the world, indigenous people carry DNA originating from more than one continent. For example, people of Amerindian descent in North and South America may also have some ancestry from Europe and Africa. When creating reference panels for the Americas and Oceania, we use only the parts of the genome with ancestry from the indigenous populations. We do this by looking at our previous assignments to select only the segments of DNA where both chromosomes have an assignment to an 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 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
Hawaii
Samoa
Tonga

For two other regions, Indigenous Cuba and Indigenous Haiti & Dominican Republic, we use windows where only one chromosome has assignment to the indigenous population. We then combine single chromosomes from two different people in the same window. This creates a window homozygous for the indigenous DNA.

2.4 Reference Panel Quality Control

For each sample, we analyze 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). Samples with large amounts of missing data are removed. We also remove samples which are likely to degrade the performance of the reference panel. Samples can be removed because 1) they are closely related to another reference sample, or 2) the underlying genetic information about a sample's origins disagrees with the sample labels, as determined through principal component analysis (PCA) (Jackson 2003, Patterson 2006) and our previous genetic analyses (Figure 2.1).

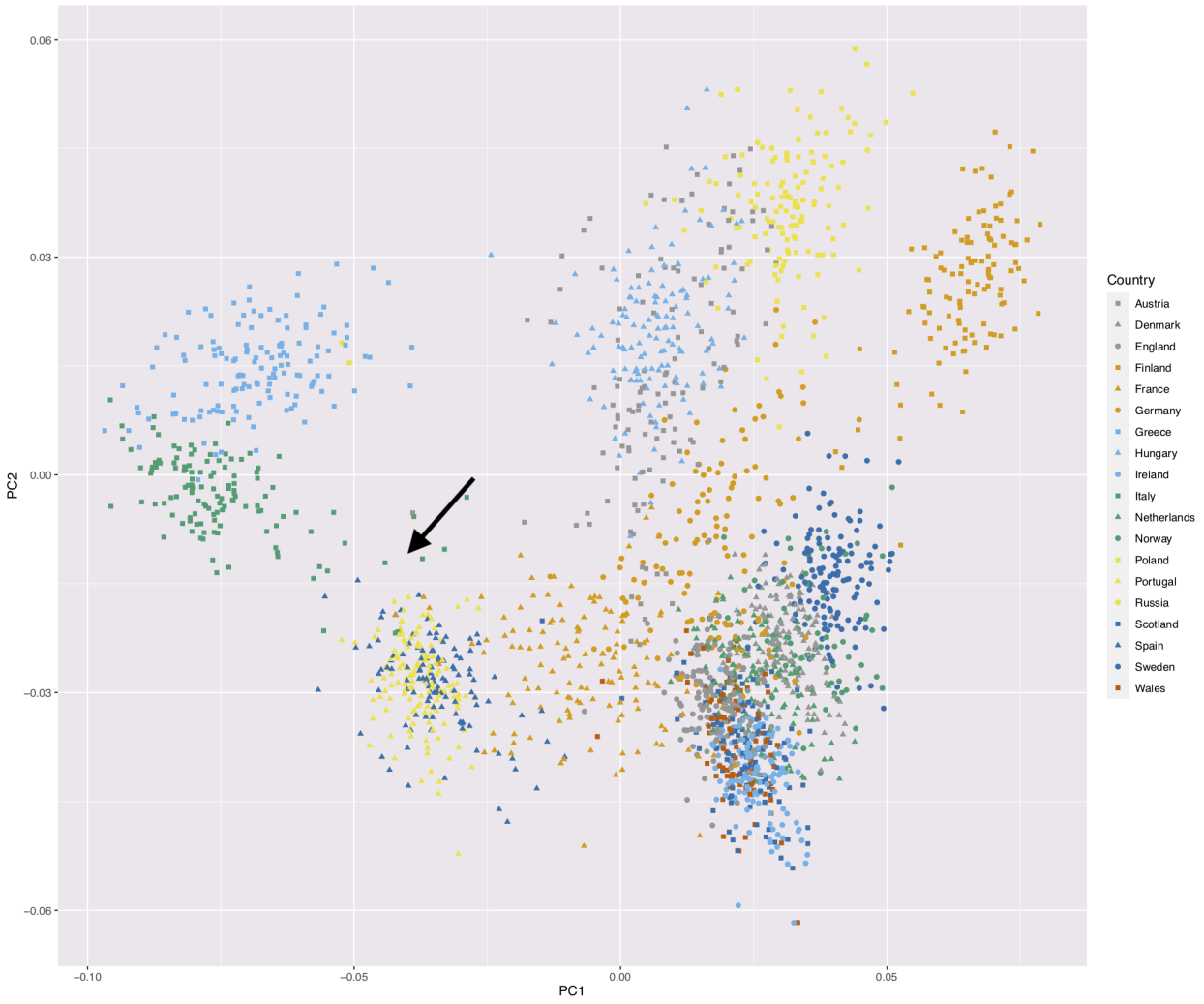

Figure 2.1: PCA Analysis on European Panel Candidates. *Scatter plot of the first two components from a principal component analysis (PCA) of candidate European samples for the AncestryDNA reference panel. Visual inspection of PCA is useful for numerous aspects of data QC. First, it can be used to identify 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.*

2.5 Updated Reference Panel

The updated AncestryDNA ancestral regions reference panel contains 116,830 samples carefully selected as described above to represent 107 global regions (Table 2.1), each with a unique genetic profile. As a comparison, our previous panel of 71,306 samples represented 88 distinct global regions.

Table 2.1: The AncestryDNA Regions Reference Panel

Region	Number of Samples
Senegal	124
Mali	336
Ivory Coast & Ghana	182
Benin & Togo	1138
Yorubaland	487
Central West Africa	497
Central Nigeria	700
North-Central Nigeria	536
Nigeria	416
Nigerian Woodlands	309
Cameroon	291
Western Bantu Peoples	184
Twa	125
Southern Bantu Peoples	148
Eastern Bantu Peoples	121
Nilotic Peoples	163
Ethiopia & Eritrea	89
Somalia	30
Khoisan, Aka & Mbuti Peoples	41
Northern Africa	636
Egypt	1281
Arabian Peninsula	1260
Levant	2000
Cyprus	1881
Anatolia & the Caucasus	2000
Iran/Persia	2000
Lower Central Asia	928
Northern Iraq & Northern Iran	1092
Burusho	17
Indo-Gangetic Plain	2000
Western Himalayas & the Hindu Kush	1048

Gujarat	998
Gulf of Khambhat	304
Southern India	262
Southwest India	718
The Deccan & the Gulf of Mannar	815
Bengal	1146
Nepal & the Himalayan Foothills	264
Tibetan Peoples	133
Northern Asia	29
Mongolia & Upper Central Asia	552
Korea	2000
Japan	136
Southern Japanese Islands	640
Northern China	245
Western China	237
Southwestern China	275
Central & Eastern China	359
Southern China	278
Dai	60
Mainland Southeast Asia	344
Maritime Southeast Asia	73
Vietnam	2000
Northern & Central Philippines	2000
Central & Southern Philippines	2000
Luzon	2000
Western Visayas	181
Guam	85
Melanesia	44
Aboriginal & Torres Strait Islander	54
Tonga	166
Samoa	112
Hawaii	363
New Zealand Maori	223
Indigenous Arctic	24

Indigenous Americas–North	1985
Indigenous Americas–Mexico	581
Indigenous Americas–Yucatan Peninsula	316
Indigenous Americas—Central	2076
Indigenous Americas—Panama & Costa Rica	466
Indigenous Cuba	9559
Indigenous Haiti & Dominican Republic	1994
Indigenous Puerto Rico	3601
Indigenous Americas–Colombia & Venezuela	3117
Indigenous Americas—Ecuador	662
Indigenous Americas–Bolivia & Peru	269
Indigenous Americas—Chile	539
Indigenous Eastern South America	2671
Ashkenazi Jews	2000
Sephardic Jews	723
Finland	2000
Sweden	2000
Denmark	1689
Norway	2000
Iceland	295
Baltics	2000
Central & Eastern Europe	2000
Russia	1590
The Balkans	2000
Eastern European Roma	852
Greece & Albania	2000
Aegean Islands	1305
Malta	1673
Sardinia	168
Southern Italy	2000
Northern Italy	2000
France	2000
Germanic Europe	2000
The Netherlands	2000

Basque	429
Spain	2000
Portugal	2000
England & Northwestern Europe	2000
Cornwall	1090
Wales	2000
Scotland	2000
Ireland	2000
Total	**116830**

3. AncestryDNA Ancestral Regions Algorithm

3.1 Algorithm Intuition and Assumptions

After establishing the reference panels, the next step is to train and tune the algorithm that infers a customer's ancestral regions by comparing nearly 300,000 selected single nucleotide polymorphisms (SNPs) from their DNA to those of the reference panel. In this comparison algorithm, we assume that an individual's DNA is a mixture of DNA from some combination of the 107 identified populations. To illustrate this concept, we show a cartoon example in Figure 3.1, where, because of recombination, a customer inherits stretches of DNA from 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 therefore the same population (for more details on DNA inheritance see our matching white paper). This means we can get more accurate results by looking at multiple nearby SNPs together as a haplotype, instead of looking at each SNP in isolation. Our algorithm takes advantage of this to greatly improve our estimates.

Our approach divides the customer's genome into 1,001 windows and 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 the reference panels for 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.

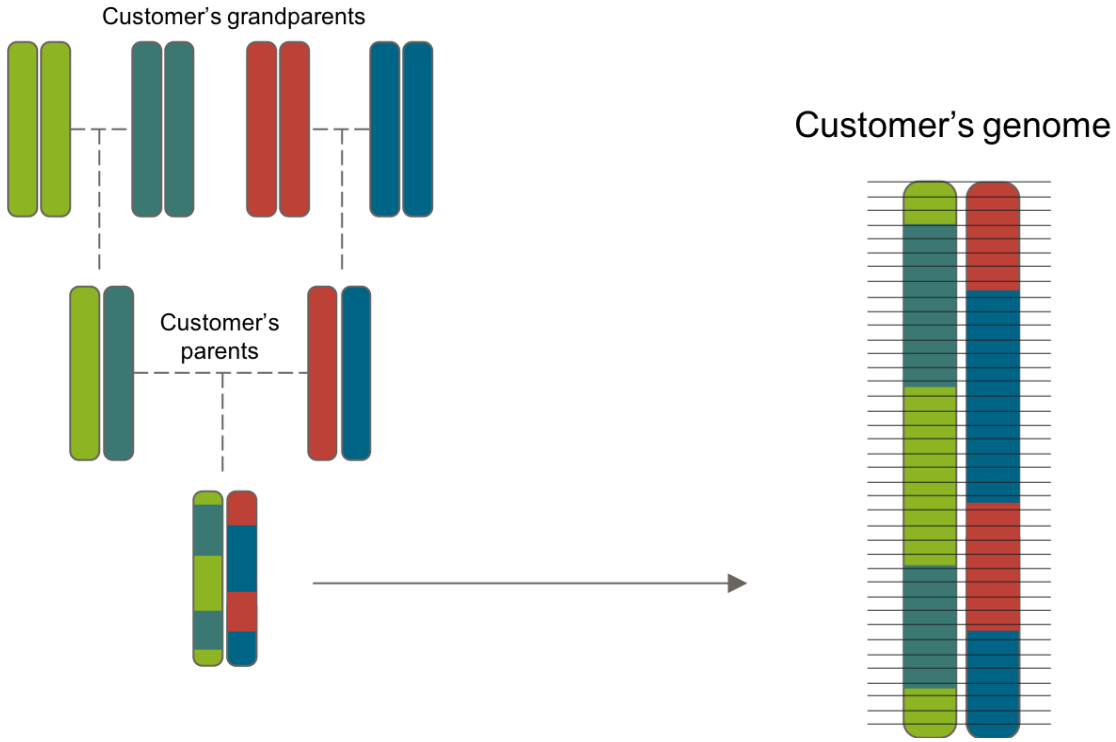

Figure 3.1: Inheritance of DNA from different populations. *On the left, we present a three-generation genetic family tree. For each individual, we show two vertical bars representing the two copies of a single chromosome present in each individual. These bars are colored to show the reference population 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 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 ancestral regions, 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. Our algorithm will assign a population to each window 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. It is important to understand that every person inherits *two* alleles, one from each parent, at each of these 700,000 sites, and that we read these sites independently. 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 region inference is to separate which letters were inherited from different parents—a process called *phasing*. Our cutting-edge technology SideView separates DNA inherited from each parent across the entire genome. Once separated, we infer the ancestral regions inherited from each parent 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 region inference to allow for incorrect phasing. In the next section, we explain how an HMM is useful in region inference, 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 each window of the genome to two of the 107 reference panels (one for each parent). A hidden Markov model is well-suited for this task because it can represent thousands of interrelated variables but still perform efficient inference—using a technique called dynamic programming—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 107 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.

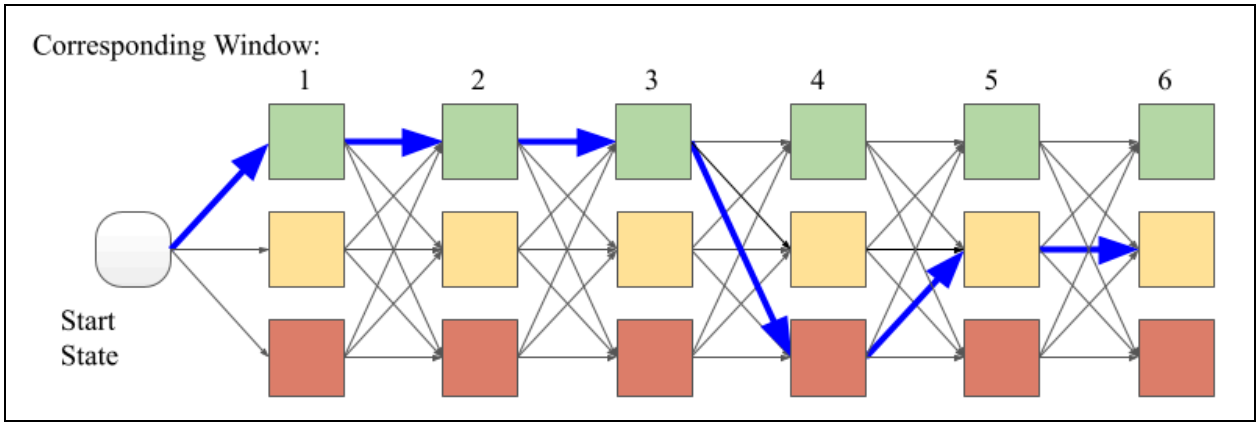

Figure 3.2: The *states* and *transitions* of an HMM representing the possible populations that explain the DNA inherited from one parent in each of several windows. *This illustration includes three populations (green, yellow, and red), and six windows. The arrows represent transitions between states, 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 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 a set of about 50,000 individuals representing diverse populations, 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. Because we are confident in the genetic separation 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 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).
IV. **Weight the emission probabilities by population.** We use results from our held-out testing data set to tune the emission probabilities so that the model produces the most accurate results possible for each population.

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 region inference, as steps I through IV 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.

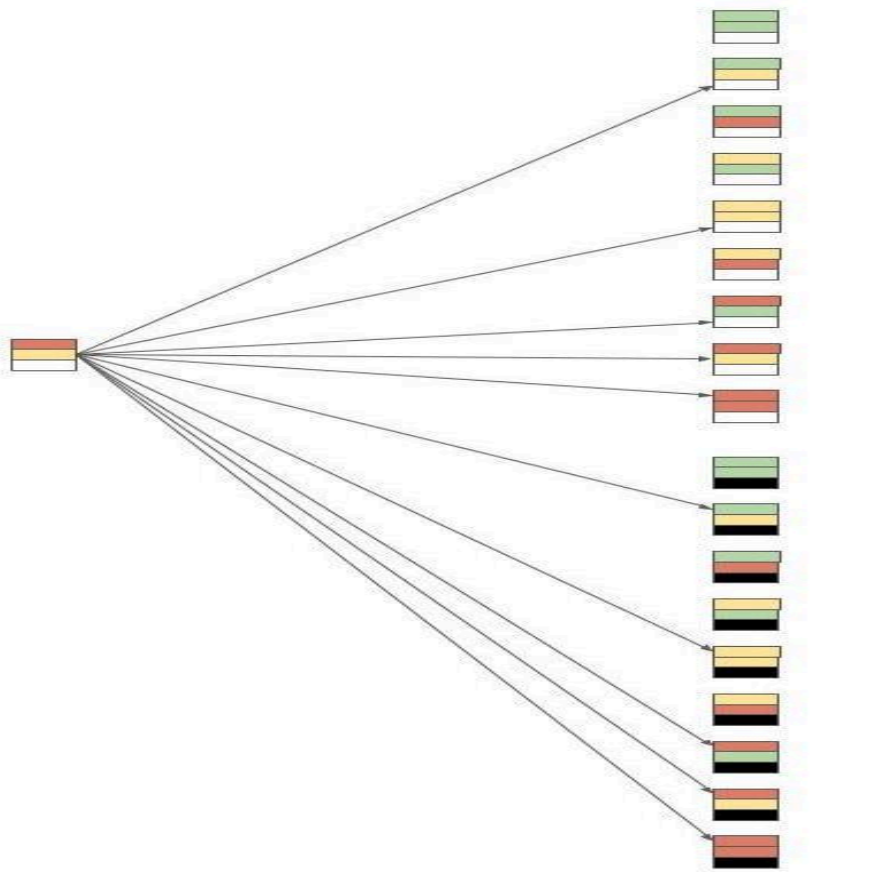

Figure 3.3: State transitions in an HMM representing *K*=3 populations. *The HMM we use in practice explains the DNA inherited from both parents simultaneously. This figure illustrates the states in a model with the same three (green, yellow, red) populations as the HMM in Figure 3.2. There are KxKx2 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 (see Section 4.5).

4. Assessing Ancestral Regions Performance

While we are developing and optimizing the estimation process, and after we finish, we repeatedly measure how well our method performs. Basically, we want to measure how close our process gets to the right answer through rigorous evaluation using a wide variety of test cases with known origins.

We use four different approaches to validate our models: 1) customer-focused simulation, 2) single-origin customers from our testing and validation sets, 3) tree-based validation, and 4) polygon creation. Each of these are described below.

4.1 Customer-focused Simulation

In any data science application, how performance is measured is the key to the algorithm's success. We use a data-centric approach to construct our testing and validation data to match the customer experience in our database.

We leverage ancestral journeys (for more information see the Genetic Communities white paper), to identify customers who share similar family histories. By aggregating 10s to 100s of thousands of family trees, we are able to establish accurate admixture patterns that differ for each group. We can then simulate separate test and evaluation data sets of genotype information, where we will also know the region results. For populations with admixture patterns that are not captured in the pedigrees, such as African American or Latin American populations, we use historical information to guide the simulations, such as in *Mooney et al.* (2023).

After analyzing the simulated data with our model, we compare the output from our model to the expected results. We measure three different statistics in aggregate and per population:

1) **Overlap** – the observed percentage for a region divided by the expected percentage for a region. Note that if the observed percentage exceeds the expected, the overlap will be above 100%.
2) **Recall** – the proportion of expected regions that are observed in the output.
3) **Precision** – the proportion of observed regions that are expected.

As we tune our models, we balance the performance of the overlap, recall, and precision statistics overall and per population. For example, as we increase the recall and overlap for one region, we often see a decrease in the precision at the same time. Our goal is to maintain as high recall as possible, while not sacrificing precision.

We note that recall and precision behave differently for very small values. Thus, we use a cut-off of 7.5% to report on performance. Expected and observed values below 7.5% have a much higher error and missing rate than those above 7.5%.

Here, we report a few numbers from a handful of our simulations from our final evaluation:

Table 4.1: Results from a simulation of 2,800 individuals reflecting histories similar to customers from Victoria Australia, London UK, Ontario Canada, Quebec Canada, New York USA, Maine USA, Pennsylvania USA, Wisconsin USA, Kentucky USA, South Carolina USA, Western States USA, Utah USA, and New Orleans Louisiana.

Region	Mean Overlap	Precision	Recall
Cornwall	52.7%	92.9%	79.6%
Denmark	87.2%	65.4%	96.0%
Central & Eastern Europe	85.7%	97.8%	97.8%
England & Northwestern Europe	106.8%	99.6%	98.5%
France	74.6%	100%	85.0%
Germanic Europe	97.4%	97.6%	96.0%
Ireland	104.2%	98.4%	99.2%
The Netherlands	76.8%	62.6%	74.0%
Norway	96.6%	96.8%	95.3%
Scotland	98.7%	97.6%	92.1%
Sweden	108.2%	95.1%	96.4%
Wales	115.0%	87.3%	96.4%
Per individual (mean)	**78.6%**	**97.2%**	**96.1%**

Table 4.2: Results from a simulation of 600 individuals reflecting histories similar to African American customers from the Southern United States, British Caribbean, and Haiti.

Region	Mean Overlap	Mean Precision	Mean Recall
Eastern Bantu Peoples	62.1%	100%	99.4%
Southern Bantu Peoples	106.5%	100%	100%
Cameroon	107.8%	99.4%	100%
Western Bantu Peoples	57.8%	99.4%	100%
England & Northwestern Europe	78.0%	100%	93.7%
Germanic Europe	77.7%	90.1%	87.0%

Ireland	118.3%	95.5%	100%
Ivory Coast & Ghana	76.0%	100%	100%
Mali	95.9%	100%	98.5%
Nigeria	97.6%	100%	99.1%
Senegal	90.2%	100%	100%
Yorubaland	117.0%	100%	100%
Per individual (mean)	**73.0%**	**99.6%**	**98.5%**

Overall, we see very strong performance across all regions, with most having precision values greater than 90% and overlap between 75% and 115%. In some regions, like *Denmark* and the *Netherlands*, we see precision values closer to 60%. We also see very strong recall and precision values across the Africa regions, suggesting that values >7.5% indicate a very confident link between a customer and that population.

We see that, on an individual average, the expected percentages overlap with 75-80% of the estimated percentages, which is a 10% improvement over the 2023 model.

4.2 Single-origin evaluation

Another way to access the performance of our model is through our evaluation dataset. For each of our reference panels, we create a testing dataset of up to 500 people to train the model weights, and a validation dataset of up to 500 people to evaluate the final model. Like the individuals used to create our reference panels, the people included in the testing and validation datasets are believed to be of a single origin, and are expected to receive 100% assignment to a specific region. We can assess each region for overlap, precision, and recall as before. After fully tuning our model, we measured the following performance metrics (regions not updated in the 2024 model are not shown):

Table 4.3: Results from 20,962 single-origin evaluation individuals.

Region	Mean Overlap	Mean Precision	Mean Recall
Western Himalayas & the Hindu Kush	67.6%	61.6%	98.9%
Cornwall	70.0%	97.8%	100%
Denmark	80.9%	63.4%	100%
Western Bantu Peoples	66.5%	49.4%	100%
Benin & Togo	91.2%	76.2%	100%
Gujarat	95.2%	67.8%	100%

Gulf of Khambhat	87.4%	75%	100%
Southern India	96.3%	36.9%	100%
Southwest India	98.5%	82.9%	100%
Lower Central Asia	95.7%	86.9%	100%
Northern Iraq & Northern Iran	99.4%	83.2%	100%
The Netherlands	69.3%	73.1%	100%
Central Nigeria	95.4%	78.0%	100%
North-Central Nigeria	90.2%	77.5%	100%
Nigeria	91.6%	71.7%	100%
Indo-Gangetic Plain	87.2%	84.9%	100%
Northern & Central Philippines	72.1%	87.7%	100%
Luzon	97.1%	52.9%	100%
Central & Southern Philippines	80.8%	78.1%	100%
Western Visayas	62.6%	38.5%	100%
Russia	84.5%	95.3%	100%
Sephardic Jews	93.2%	100%	100%
The Deccan & the Gulf of Mannar	97.0%	56.1%	100%
Twa	98.9%	100%	100%
Yorubaland	89.1%	89.6%	100%
Central West Africa	88.7%	71.1%	100%
Aegean Islands	85.8%	94.6%	100%
Anatolia & the Caucasus	90.3%	70.0%	100%
Arabian Peninsula	85.0%	93.2%	99.7%
Baltics	96.6%	72.8%	100%
Basque	97.0%	39.8%	100%
Bangladesh	93.7%	77.7%	100%
Burusho	34.5%	100%	100%
Cameroon	84.4%	56.7%	100%
Cyprus	98.0%	98.8%	100%
Dai	64.9%	55.6%	100%
Eastern Bantu Peoples	76.3%	70.7%	100%
Eastern Europe	88.2%	43.6%	100%
Eastern European Roma	86.0%	100%	100%
Egypt	96.0%	97.6%	100%
England & Northwestern Europe	79.7%	34.5%	100%
Ethiopia & Eritrea	93.5%	38.2%	100%
Finland	96.9%	77.5%	100%

France	74.8%	91.8%	100%
Germanic Europe	73.6%	32.8%	100%
Greece & Albania	83.2%	88.0%	100%
Guam	68.0%	100%	100%
Iceland	97.0%	97.3%	100%
Iran/Persia	85.7%	84.6%	100%
Ireland	96.8%	51.4%	100%
Ivory Coast & Ghana	61.1%	61.1%	100%
Japan	95.7%	32.4%	100%
Ashkenazi Jews	99.1%	99.2%	100%
Khoisan, Aka & Mbuti Peoples	96.0%	56.3%	100%
Korea	97.9%	94.0%	100%
Levant	97.5%	78.0%	100%
Mainland Southeast Asia	79.8%	94.5%	100%
Mali	94.7%	84.8%	100%
Malta	97.2%	100%	100%
Maritime Southeast Asia	53.9%	81.0%	100%
Melanesia	98.1%	84.6%	100%
Mongolia & Upper Central Asia	86.3%	98.6%	100%
Nepal & the Himalayan Foothills	89.9%	100%	100%
Nigerian Woodlands	92.5%	98.7%	100%
Nilotic Peoples	76.7%	74.1%	100%
Northern Africa	96.7%	96.9%	100%
Northern Asia	51.1%	85.7%	100%
Northern Italy	81.1%	98.9%	100%
Norway	87.0%	82.1%	100%
Portugal	93.8%	77.6%	100%
Sardinia	98.8%	100%	100%
Scotland	69.3%	61.0%	100%
Senegal	95.6%	85.7%	100%
Somalia	81.1%	77.8%	100%
Southern Bantu Peoples	96.4%	82.9%	100%
Southern Italy	96.4%	54.5%	100%
Southern Japanese Islands	85.7%	97.0%	100%
Spain	85.6%	70.8%	100%
Sweden	86.9%	60.4%	100%
The Balkans	70.4%	67.0%	100%

Tibetan Peoples	97.9%	60.0%	100%
Vietnam	94.8%	94.7%	100%
Wales	93.1%	83.8%	100%
Per individual (mean)	**87.9%**	**82.5%**	**100%**

For some regions, such as *Burusho* and *Northern Asia*, the numbers are not as high, with average assignments of 34% and 51% to the correct region, respectively. However, even if the prediction accuracy falls short of 100% for some regions, the remaining percentage is still assigned to nearby regions. For example, individuals from *Burusho* might get some assignments to the *Indo-Gangetic Plain* region, and individuals from *Northern Asia* might get some level of assignment to *Mongolia & Upper Central Asia*. Similarly, when precision is low, we find that it is people from nearby-regions that get the extra assignment.

We found that the majority of our regions have precision and overlap that are both greater than 75%. Additionally, we found that the average overlap for single-origin people is 87.9%, which is a significant improvement from 2023 (greater than 12%).

In summary, although our models are tuned based on admixed samples, we are pleased to report an overall improvement in 2024 in accuracy for people of single-origin.

4.3 Tree-based validation

An independent way to validate our model is to look at the ancestral regions results of people with deep genealogical roots back to the same country or part of a country. 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 all 481 regions of the world that we considered. For example, Figure 4.1 shows the average assignments for 200 customers with all four grandparents (or older) born in Germany (top) and 200 customers with all four grandparents born in South Korea (bottom). As you can see, while most of

their assignment is to the expected corresponding regions, *Germanic Europe* and *Korea*, other regions do appear in small but significant amounts. These analyses help ensure that results for people from a geographic area agree with expectations.

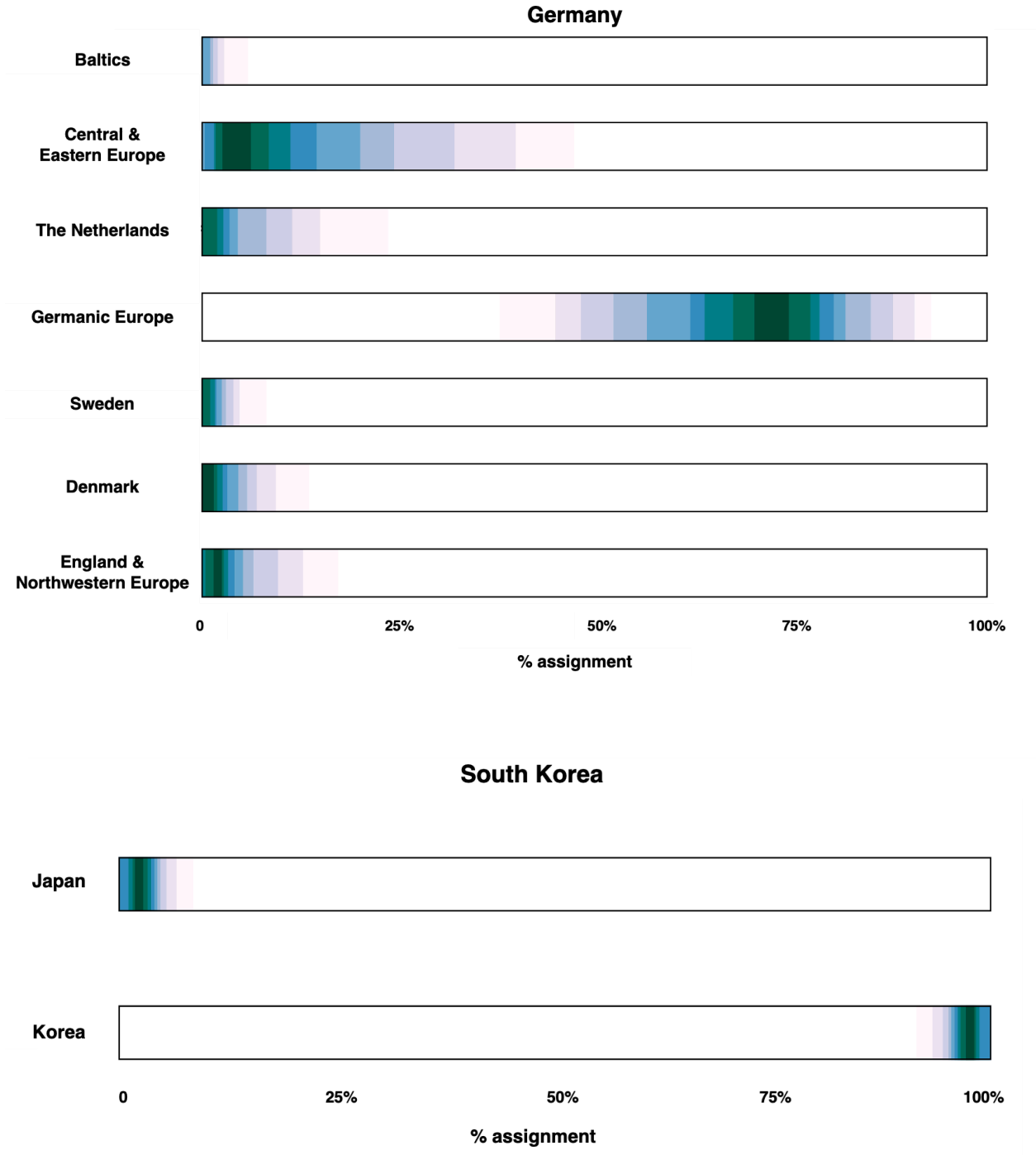

Figure 4.1 Average assignments based on grandparents' birth location. *Region assignment distribution for customers with all*

four grandparents born in the same country. Germany (top), and South Korea (bottom). Dark green is the middle 50th percentile, with the distribution bucketed and colored by percentile.

4.4 Regional Polygon Construction

The process we use to create polygons for each of our 107 regions also helps to validate our model. 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 results and geographic information used to define the polygon for our England & Northwestern Europe region.

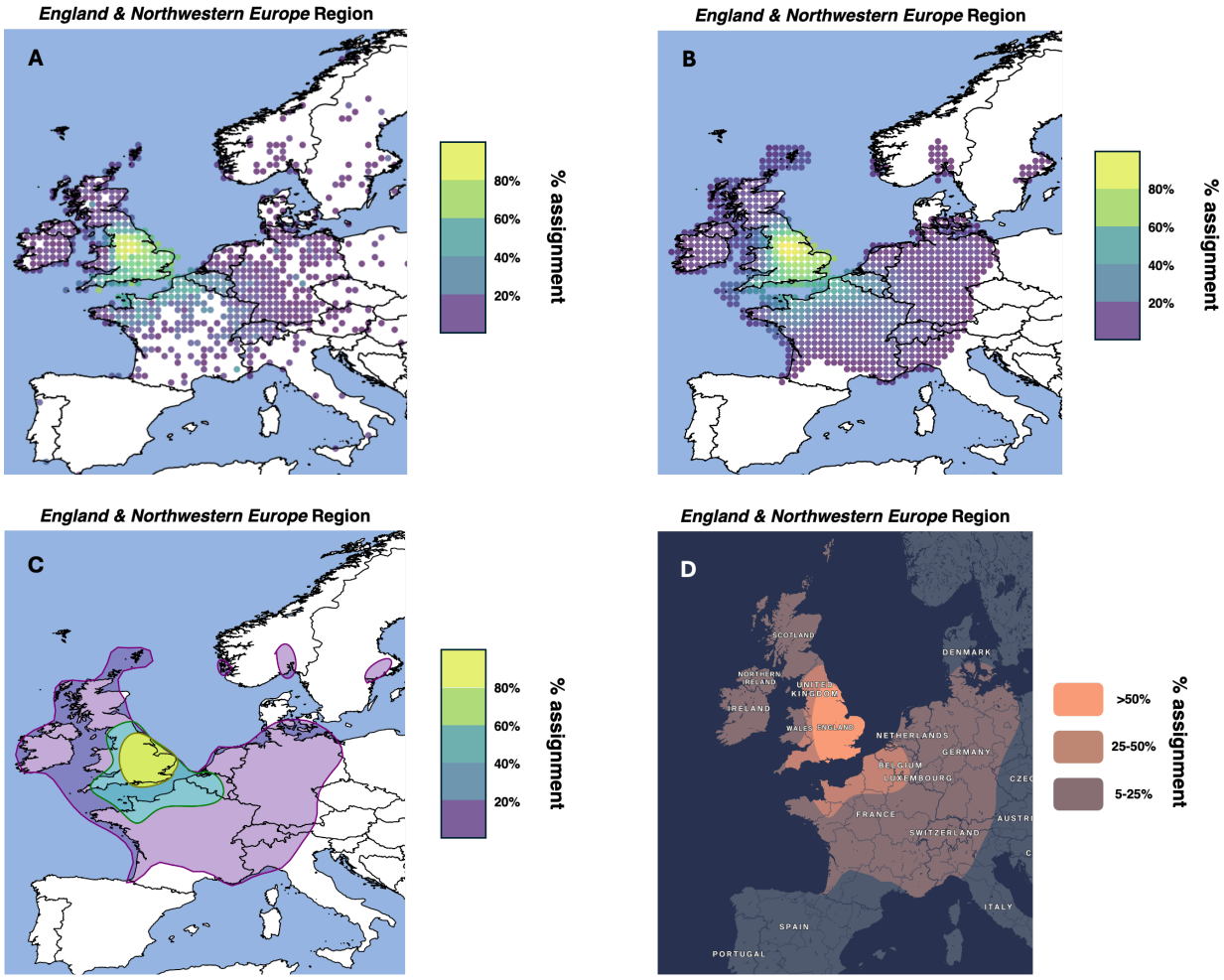

Figure 4.2: Using geographic sample locations to draw regional polygons. *Panel A shows the distribution of the England & Northwestern Europe region predicted for a set of samples with geographic information. Samples are assigned to grids of 0.5 degrees longitude by 0.5 degrees latitude based on the average birth location of their grandparents. The color of each grid squarepoint on the map represents the average percentage of England & Northwestern Europe for samples from each grid. Panel B*

shows the maps after filling in missing grids using an imputation method and smoothing the results. Panel C shows the information processed with further smoothing, creating the outlines representing the ancestry regions shown to customers. Panel D shows the final polygon presented in a customer's results.

In Figure 4.2A, we show the amount of our *England & Northwestern Europe* region assigned to a combination of reference panel evaluation samples and customers with deep roots from the same country. Figure 4.2B shows the results after imputing values to fill in gaps in our map grid and applying smoothing methods to make the plot less spotty. It is clear from the plot that there is a gradient of assignment in this area that is centered in England and quickly tapers off in surrounding areas. For example, the next level of concentration, represented by light blue in the image, is in northeastern France and Belgium. The gradient continues to diminish as represented in purple, with the borders reaching as far away as Northern Italy, Norway, and Switzerland.

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. Additionally, polygons representing some of our regions have hand-drawn components to describe minority populations that may not be explicitly defined by geography. For regions which are data driven, these polygons are a powerful tool that we use to validate each one of our regions.

4.5 Reporting uncertainty of estimated values

As mentioned in Section 3.5, we report a range for each ancestral region that we deliver to customers. For example, we might report someone as 40% *England & Northwestern Europe* with a range of 30-60%. This means that the model reports the most likely estimate of 40% *England & Northwestern Europe*, but that our model also supports an estimate anywhere between 30% and 60% *England & Northwestern Europe*. We run a separate analysis to validate the range results in our simulation datasets to ensure that the expected value is captured within the range most of the time.

5. Finer Scale Subregion Inference

Algorithms that infer your ancestral origins, like the one described above, are designed to assign portions of your DNA to various global populations based on similarity to people in a representative reference panel.

However, even specific and commonly assigned regions, like *England and Northwestern Europe*, can be frustratingly broad to customers and family genealogists when they are looking for records in a region to further their family history research and discovery.

In order to provide even greater resolution and specificity for ancestral origins, we developed a complementary method for connecting customers to people and places around the world—called "subregions". This new feature uses distinct subregion reference panels which have been created from individuals with deep genealogical roots in a specific geographic area. We then identify short segments of matching DNA that a customer shares with the people in the subregion reference panel. The method leverages the principle from population genetics that people from the same population share more DNA with each other than with people from other populations. Additionally, by focusing on shorter segments of matching DNA, we are more likely to detect more distant (i.e., older) genetic connections. These results can indicate a genetic connection to the people of a specific geographic area in the past several hundred years. Overall, subregion assignments provide an orthogonal set of evidence to help customers narrow down their search for family origins.

5.1 Construct subregion reference panels

To develop reference panels for subregions, we identify individuals who can represent more specific geographic areas. This is made possible by Ancestry's enormous database of genetic samples paired with genealogical trees provided by members. The subregion reference panel candidates have multigenerational family trees where all the ancestors are from a shared and specific geographic area. For example, *Connacht* is one of the subregions contained in our *Ireland* ancestral region. We developed the *Connacht* subregion reference panel using individuals who had multigenerational trees linked to their DNA tests, and where the earliest ancestors (i.e., terminal nodes) in those trees were all born in Connacht, Ireland. By combining genetic information with information from trees, we can build separate reference panels that represent people with deep ancestral roots to a geographic area.

Our subregion method leverages the principle that people from the same population share more DNA with each other than they do with other populations. This is most often true over large geographic distances, for example comparing a subregion in Ireland with a subregion in Bulgaria. However, populations that are closer in proximity may be more difficult to differentiate using genetics. This could lead to inaccurate subregion results for customers.

To address this, we rigorously tested the **precision** ($\frac{true\ positives}{(\ true\ positives + false\ positives\ )}$) and **recall** ($\frac{true\ positives}{(\ true\ positives + false\ negatives\ )}$) for each of our subregion reference panels. In cases where we were able to maintain a finer level of granularity to our subregions, we did so. In situations where assignments were inconsistent or inaccurate, we merged together neighboring subregions to improve performance. For example, in testing the *Kosovo* and *Montenegro* subregion reference panels, we found many people with trees tracing back to Kosovo were instead assigned to Montenegro, and vice versa. By combining these reference panels into a single *Kosovo & Montenegro* subregion we recovered a higher level of performance. Overtime, as more individuals take tests and upload trees, it may be possible to revisit these reference panels and add greater granularity to some subregions.

5.2 Calculate matching to subregion

As DNA is passed down from an initial founder to subsequent descendants, recombination breaks up the inherited matching DNA segments. As a result, when searching for deeper ancestral connections, a living individual is likely to share mostly short segments of matching DNA with the population of their ancestral origins. Using our matching algorithm (see our white paper for details) we are able to confidently detect short matching segments of DNA shared by a customer and any individuals in our subregion reference panels. This gives us the opportunity to connect members with people and geographic areas that are part of their more distant family history.

However, while short segments can be informative about deeper ancestral connections, it is also the case that shorter segments of matching DNA between people are more likely to be the result of chance than the result of a shared common ancestor. This significantly complicates the process of connecting an individual to their ancestral subregion just by looking for matching DNA segments in a reference panel.

We therefore developed an assignment method that makes use of both a match score and a subregion specific score threshold. The match score is calculated per individual as the average amount of DNA (in centimorgans) shared with the subregion reference panel, selecting only the top matching reference panel samples. We then determined a match score threshold for each subregion by evaluating precision and recall metrics. Individuals whose amount of shared DNA with the reference panel exceeded the subregion's specific threshold were assigned that subregion.

5.3 Evaluating score thresholds for subregion performance

Each population represented by a subregion has its own unique history, impacting features like the level of genetic similarity within the population and similarity to other populations. Additionally, the subregion populations are represented to a different degree in our customer base. This means the expected amount of matching between the subregion reference panel and a person with deep ancestral roots in that area will vary for each subregion. As a result, we needed to develop subregion specific thresholds for our match scores to confidently assign a customer to a subregion.

Determining this threshold, however, is complicated by the fact that for most customers, we don't have a ground-truth for what subregion assignment to expect. Many customers have either an incomplete tree, short tree, or no tree from which we can evaluate their expected subregion assignment. Those individuals who do have high quality trees against which we could test our assignment were likely already included in our reference panels.

We therefore adopted a simulation-based approach, leveraging our customer database and those individuals whose pedigrees gave us confidence in their subregion assignment. We selected individuals we were confident belonged to specific subregions and used them as founders of a simulated pedigree. The simulation involved successive generations of admixture in the descendants of the founders.These simulations allowed us to distribute the genetic signal of a specific subregion among descendants, and still be able to trace the correct subregion assignment. Using these simulated descendents, we evaluated precision and recall when applying various match score thresholds for each subregion. To provide the greatest insight to customers, we established three threshold levels based on increasing precision values of 60%, 75%, and 90%. These levels indicated a "moderate", "strong", and "very strong" genetic connection to the subregion, respectively.

6. Future Refinement

While AncestryDNA is extremely proud of the updates in this release, we plan 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.

Each new release of genetic ancestral regions 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.

7. References

- AncestryDNA White Papers: https://support.ancestry.com/s/article/AncestryDNA-White-Papers
- 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.
- 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.
- Mooney JA, Agranat-Tamir L, Pritchard JK, Rosenberg NA. On the number of genealogical ancestors tracing to the source groups of an admixed population. Genetics. 2023; 224(3), iyad079.
- 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.