

Ancestral Regions 2026 White Paper

Last updated October 08, 2026

Rekha Angara, Jeffrey Adrion, Ross Curtis, Jenna Lang, Keith Noto, Richard Olpin, Genivaldo Silva, Natalie Swinford Tibi, Yong Wang, Aaron Wolf, Ju Zhang (in alphabetical order)

1. 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 its [Genetic Communities](#) technology, 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 146 regions to 302). 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 302 reference panels for 2026
2. How our updated region assignment algorithm works to estimate a customer’s genetic ancestral 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 origins 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 common 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.

Parental haplotype — The unbroken sequence of genetic markers inherited together on a single chromosome from one specific parent.

Population — A group of individuals who share distinctive DNA patterns as a result of a common geographic origin, shared history, or cultural ties over generations.

Phasing — The assignment of DNA to contiguous segments corresponding to the DNA inherited from each biological parent. 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 a person from Korea will have 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 302 regions in our reference panels. 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, in part, 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 29 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**) 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 more than 4,500 DNA-based communities we have discovered as the bases for reference panels.

Our Genetic Communities technology identifies networks of Ancestry DNA test takers that are highly interconnected due to sharing DNA from a common set of ancestors (Figure 2.1). We are also able to look at family tree information in aggregate for these networks to identify shared places of origin, patterns of movement, statistically enriched surnames, and genetic ancestral region results. These data provide insight into the specific identity and story of these genetically related groups.

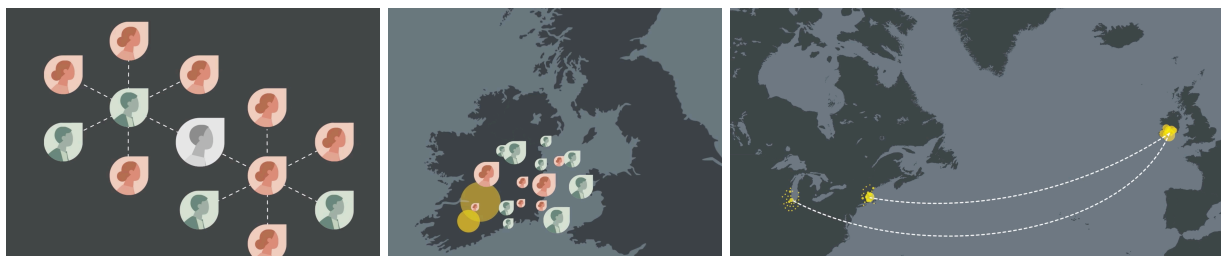


Figure 2.1: Genetic Communities technology identifies networks of interconnected Ancestry members who share a common history. We analyze our match database of 29 million DNA test takers and identify networks of connected matches. We then use information from these members' trees to identify common places of origin and patterns of movement that give indications about the identity and shared history of these genetic families.

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. Identifying individuals to represent a population based on patterns of genetic similarity between these individuals is preferable to hand-curating a set of individuals 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 complicating factor is that individuals' membership in 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 and do not have recently admixed family origins. We did this by filtering out individuals who had a weak genetic connection to their assigned networks and who were assigned to multiple networks from different populations.
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 regions 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 DNA. 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 remove individuals who showed a low level of genetic connectedness to the specific region and a high level of genetic connectedness to outside regions as measured by their number of matches to those networks.
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 filtering approach based on IBD-sharing 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 several thousand individuals per reference panel to train our model ("training

set”), along with several hundred testing samples to tune parameters (“test set”), and another several hundred 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:

- Samples from worldwide populations from a public project called the Human Genome Diversity Project (HGDP) (Cann *et al.* 2002; Cavalli-Sforza 2005)
- Samples from the 1000 Genomes Project (McVean *et al.*, 2012)
- Samples 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 origins. We then use aggregated meta-data (e.g., self-reported ethnicity, spoken language, common surnames) to identify the groups. We sample a subset of the total individuals to include in our reference panel.

2.4 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 DNA from Europe and Africa. When creating reference panels for the Americas, we use only the parts of the genome inherited from the indigenous populations. We do this by looking at our previous assignments to select only the segments of DNA where one or both copies of the chromosomes have an assignment to an indigenous population and combining these segments across individuals. So, whereas most of our ancestral 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 51 regions where we employ this approach are:

Pampas & Patagonia (Argentina)
Río de la Plata Basin (Argentina & Uruguay)
Coastal Belize

Altiplano & Gran Chaco (Bolivia)
Mata Atlântica (Brazil)
Sertão & Amazon Basin (Brazil)
Atlantic Pampas & Highlands (Brazil & Uruguay)
Northern Prairies & Great Lakes (Canada & USA)
Atlantic Northeast (Canada & USA)
Southern Andean Coast (Chile)
North Andean Coast (Colombia)
High Andean Heartland (Colombia)
The Pacific Chocó (Colombia)
Talamanca Highlands & Nicoya Coastlands
Cubanacán (Cuba)
Quisqueya (Dominican Republic)
Equatorial Andes & Pacific Coast (Ecuador)
Cuzcatlán (El Salvador)
Mayan Highlands (Guatemala)
Honduran Highlands & Caribbean Coast
Lenca Heartland & Pacific Slopes (Honduras & El Salvador)
Centro-Occidente—Norte (Mexico)
Baja California Peninsula (Mexico)
Chihuahuan Desert & Sierras (Mexico)
Centro-Occidente—Sur (Mexico)
Western Sierra Madre Highlands (Mexico)
Eastern Sierra Madre & Gulf Coast (Mexico)
Centro-Oriente Highlands (Mexico)
El Bajío (Mexico)
Jalisco Highlands (Mexico)
Noreste & Rio Grande (Mexico & USA)
Huasteca & Gulf Lowlands (Mexico)
Gulf of California (Mexico)
Sonoran Desert & Gulf Coast (Mexico & USA)
Southern Sierra & Tehuantepec (Mexico)
Yucatán & Petén Basin (Mexico)
Nicaragua

Isthmus of Panama
Gran Chaco (Paraguay)
Central Andean Altiplano (Peru)
Northern Andean Slopes (Peru)
Borikén (Puerto Rico)
Southern Rockies & Upper Rio Grande (USA)
Great Plains (USA)
Northern Rio Grande Basin (USA)
North Pacific Coast (Canada & USA)
Gulf Coast & Mississippi Basin (USA)
Colorado Plateau (USA)
Coastal Cordillera & Sistema Coriano (Venezuela)
Orinoco Delta (Venezuela)
Maracaibo Basin (Venezuela)

2.5 Reference Panel Quality Control

For each sample, we analyze a set of approximately 400,000 SNPs that are shared between the Illumina OmniExpress platform and the Illumina HumanHap 650Y platform (which was used to genotype HGDGP samples). Samples with large amounts of missing data are removed.

2.6 Updated Reference Panel

The updated AncestryDNA ancestral regions reference panel contains over 295,000 samples carefully selected as described above to represent 302 global regions (Table 2.1), each with a unique genetic profile. As a comparison, our previous panel of 185,000 samples represented 146 distinct global regions. This represents an almost 60% growth in the size of our reference panel and a doubling in the number of distinct global regions for our 2026 regions update.

Table 2.1: The List of AncestryDNA Regions Included in the 2026 Update

Ancestral Regions								
Connacht, Ireland	Canary Islands	Western Carpathian Basin	Central China	Uttar Pradesh & Bihar	Jordan Valley & West Bank	Benin	El Bajío (Mexico)	Atlantic Pampas & Highlands (Brazil & Uruguay)
Donegal, Ireland	Northern Spain	Southern Poland	Pearl River & Lingnan Coast	Northern Punjab	Middle Euphrates	Southern Benin—Yoruba Peoples	Jalisco Highlands (Mexico)	Gran Chaco (Paraguay)
Leinster, Ireland	Spain	Lesser Poland	Guangdong, China	Northeastern Punjab	Lebanon & Northern Syria	Benin—Western Gbe Peoples	Noreste & Rio Grande (Mexico & USA)	Guam
Munster, Ireland	Basque	Northern Poland	Indigenous Siberia	Central Punjab	Northern Levant	Central Nigeria—Tiv People	Huasteca & Gulf Lowlands (Mexico)	Fiji
Ulster, Ireland	Malta	Central & Eastern Poland	Northwestern Mongolia	Khyber Pakhtunkhwa & Gilgit-Baltistan	Lebanon & Southern Syria	Nigeria—Jos Plateau	Gulf of California (Mexico)	Papua

Isle of Man	Sardinia	Eastern Slovakia	Northern Mongolia & Southern Siberia—Buryat	Northern India & Pakistan	Cyprus	Nigeria—Middle Belt	Sonoran Desert & Gulf Coast (Mexico & USA)	Vanuatu
North East Scotland	Central Italy	Carpathian Basin & Uplands	Western Mongolia	Western Kashmir	Armenia	Southern Kaduna & Central Nigeria Plateau	Southern Sierra & Tehuantepec (Mexico)	Eastern Melanesia
Hebrides & Western Highlands, Scotland	Northeastern Italy	East-Central European Lowlands	Kyrgyzstan	Pashtunkhwa	Eastern Anatolia Armenians	Central Nigeria—Gbagyi People	Southern Rockies & Upper Rio Grande (USA)	Western Melanesia
Scottish Lowlands & Central Belt	Northwestern Italy	Western Ukraine	Central Asia	Western Punjab	Cilicia Armenians	Central Nigeria—Migili People	Northern Rio Grande Basin (USA)	Aboriginal and/or Torres Strait Islander Peoples
Southern Wales	Southern Italy	Eastern Czechia	Hazara	Punjab & Sindh	Central Turkey	Yorubaland	Yucatán & Petén Basin (Mexico)	Tonga
Brittany	Sicily	Western Czechia	Central Mongolia	Sindh	Eastern Turkey	Igboland	Coastal Belize	Samoa
Northern Wales & North West England	Crete	Russia	Northwestern Mongolia & Siberia—Turkic Peoples	Bengal	Southeastern Turkey	Southwestern Cameroon	Nicaragua	Hawai'i
Devon & Somerset	Ionian Islands	Maritime Southeast Asia	Andhra Pradesh	Northeastern Bangladesh & Assam	South Caucasus	Cameroon—Nso & Highland Peoples	Honduran Highlands & Caribbean Coast	Tahiti
East Midlands & Yorkshire	Northern & Central Greece	Central & Southern Vietnam	Kerala	Nepal & the Himalayan Foothills	Georgia	Cameroon—Ring Kingdom Peoples	Lenca Heartland & Pacific Slopes (Honduras & El Salvador)	New Zealand Māori
Northern England	Southern Greece	Northern Vietnam	Central Kerala	Bhutan	Western Turkey	Cameroon—Widikum Peoples	Mayan Highlands (Guatemala)	
Southeastern England	Pontian Greeks	Northern Luzon	Tamil Nadu	Central & Eastern Nepal	Northern Turkey	Cameroon—Bamenda Highlands Peoples	Cuzcatlán (El Salvador)	
West Midlands	Aegean Islands	Southern Luzon	Tamil Nadu—Iyengar	West Bengal & Northern India	Burundi & Rwanda	Cameroon—Bamileke Plateau	Talamanca Highlands & Nicoya Coastlands	
Cornwall	Northwestern Croatia & Istria	Central Luzon	Telangana & Andhra Pradesh	Atlantic Sahara	Uganda—Central Sudanic Peoples	Cameroon—Bangwa & Bamileke Peoples	Isthmus of Panama	
Germans in Russia	Central Croatia	Bicol	Kerala—Varghese	Northern Morocco & Mediterranean Coast	Western Nilotic Peoples	Western Bantu & Forest Peoples	Cubanacán (Cuba)	
Northwestern Germany	Western Balkans & Sava Plains	Western Visayas	Sri Lanka	Central & Eastern Maghreb	Kenya—Eastern Nilotic Peoples	DRC—Luba Peoples	Quisqueya (Dominican Republic)	
Bavaria & Western Austria	Central Balkans	Central Visayas	Indian Ocean—Tamil	Northern Saharan Oases	Kenya & Uganda—Eastern Bantu Peoples	Southern Bantu Peoples	Borikén (Puerto Rico)	
Northeastern Germany	Dalmatia	Visayas & Mindanao	Andhra Pradesh & Southern Karnataka	Upper Egypt	Ethiopia	Zimbabwe	North Andean Coast (Colombia)	
The Netherlands	Southwestern Balkans	Eastern Visayas	Gujarat	Middle Egypt	Eritrea	Alaska-Yukon (Canada & USA)	High Andean Heartland (Colombia)	
France	Romania	Tai Peoples	Parsis	Egypt	Somalia—Equatorial Coast	Northern Prairies & Great Lakes (Canada & USA)	The Pacific Chocó (Colombia)	
Belgium	Slovenia	Cambodia	South Gujarat	Sudan	Horn of Africa	Atlantic Northeast (Canada & USA)	Coastal Cordillera & Sistema Coriano (Venezuela)	
Switzerland	Albania	Myanmar	Gulf of Khambhat	Arabian Peninsula	Uganda Batwa	Great Plains (USA)	Orinoco Delta (Venezuela)	
Acadia	Bulgaria	Mainland Southeast Asia—Hmong People	Goa	Northern Iraq & Iran—Assyrian Peoples	Senegal	North Pacific Coast (Canada & USA)	Maracaibo Basin (Venezuela)	
Quebec	Eastern European Roma	Mainland Southeast Asia—Iu Mien People	Gujarat—Darji	Northern Iraq	Southern Senegal & Guinea-Bissau	Gulf Coast & Mississippi Basin (USA)	Equatorial Andes & Pacific Coast (Ecuador)	
Denmark	Sephardic Jews in the Eastern Mediterranean	Japan	Maharashtra	Southern Iraq	Hausa & Fulani Peoples	Colorado Plateau (USA)	Central Andean Altiplano (Peru)	
Sweden	Sephardic Jews in Northern Africa	Southern Japanese Islands	Mumbai	Central Iran	Liberia	Centro-Occidente—Norte (Mexico)	Northern Andean Slopes (Peru)	
Norway	Ashkenazi Jews in Eastern Europe & Russia	Chūgoku Chihō, Japan	Coastal Karnataka	Armenians in Iran	Sierra Leone	Baja California Peninsula (Mexico)	Altiplano & Gran Chaco (Bolivia)	
Iceland	Ashkenazi Jews in Central & Southeastern Europe	Korea	Eastern Punjab	Iranian Azerbaijan	Ivory Coast & Southwestern Mali	Chihuahuan Desert & Sierras (Mexico)	Southern Andean Coast (Chile)	
Finland	Eastern Mizrahi	Eastern & Northeastern China	Kashmir & Punjab Foothills	Southern Iran	Mali—Mandé Peoples	Centro-Occidente—Sur (Mexico)	Pampas & Patagonia (Argentina)	
Azores	Western Mizrahi	Northern China	Northwestern India	Iran	Mali—Dogon Peoples	Western Sierra Madre Highlands (Mexico)	Río de la Plata Basin (Argentina & Uruguay)	
Madeira	Estonia & Latvia	Southwestern China	Rajasthan & Gujarat, India	Lebanon	Ghana & Western Togo	Eastern Sierra Madre & Gulf Coast (Mexico)	Mata Atlântica (Brazil)	
Portugal	Lithuania	Taiwan & Southeastern China	Jammu & Kashmir	Southern Levant	Benin—Eastern Gbe Peoples	Centro-Oriente Highlands (Mexico)	Sertão & Amazon Basin (Brazil)	

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 region assignment algorithm that infers a customer's ancestral regions by comparing over 400,000 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 302 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.

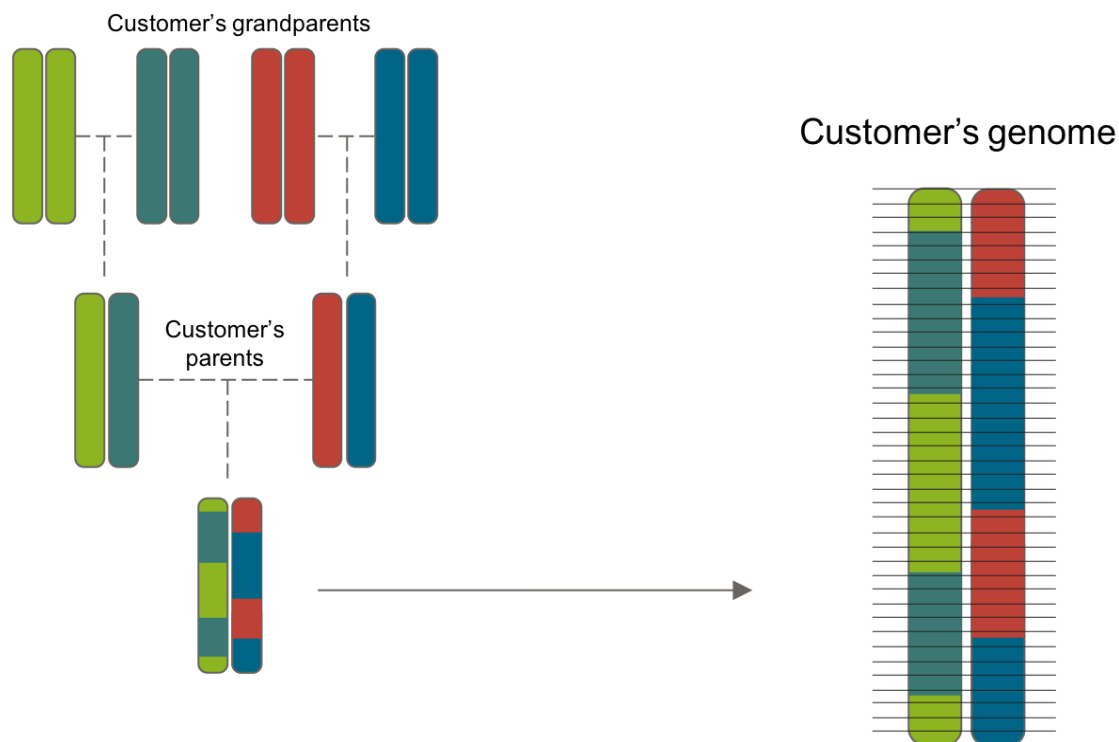


Figure 3.1: Inheritance of DNA from different populations. On the left, we present a three-generation genetic family tree. For each individual, we show two vertical bars representing the two copies of a single chromosome present in each individual. These bars are colored to show the reference population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

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 a large number of 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). Determining the ancestral origin of a piece of DNA follows two main steps:

1. **Initial Inference:** Determine the ancestral origin of every section of the customer's DNA independently by comparison to our reference panel.
2. **Smoothing:** Refine that initial inference to improve accuracy using an AI transformer architecture.

Once we have determined the most likely ancestral origin for each window of a customer's DNA, we combine information from all the windows to estimate what overall portion of the customer's genome came from each population.

The region assignment algorithm we have developed for our 2026 Regions update differs significantly from the algorithm we have used in previous updates. The differences impact both how we perform our initial inference and our smoothing steps. In previous releases we utilized a haplotype cluster model in which we studied our reference panel to identify DNA patterns that were commonly shared among members of specific regions and less common among other regions. This approach was robust and efficient, since we could build the haplotype cluster model once and then we could compare any customer's DNA to the model to label their DNA windows with the most likely ancestral regions of origin.

In our updated region assignment algorithm for 2026, we now compare a customer's DNA in a window directly against our reference panel members' DNA in the same window. This direct comparison to our reference panel allows us to identify less common and more subtle DNA patterns that are indicative of the ancestral origins for the DNA in that window.

Additionally, in past releases, our region assignment algorithm relied on a hidden Markov model for identifying the most likely sequences of regions across a customer's DNA. The nature of this model allowed us to adjust it to be more or less permissive of transitions from one region assignment to another across windows of a customer's DNA. This way, we could adjust our algorithm to improve our overall accuracy by "smoothing" noise in the region assignment process.

Our regions assignment algorithm for 2026 does not use a hidden Markov model, and instead relies on an AI transformer architecture to smooth region assignments across windows and provide a high level of overall accuracy for customer's results. The transformer does this by simultaneously drawing on

information from a large number of surrounding windows in the customer's DNA to provide improved predictions for each individual window.

Sections 3.3 through 3.7 go into more detail about these new components of our region assignment algorithm.

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 400,000 SNPs that are shared with all members of the reference panel.

SideView uses DNA shared with 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, our region assignment algorithm includes a step that detects and corrects places where the separation of the two parental copies has been swapped. Section 3.5 describes it.

3.3 Comparing Each Window to the Reference Panel for Initial Inference of Region Assignments

Our region assignment algorithm begins with dividing the customer's genome into a large number of windows, and separating the DNA in each of these windows into the separate portions inherited from each parent—the **parental haplotypes**. For each window and each parental haplotype we calculate how likely it is that the piece of DNA is inherited from any of our 302 ancestral regions. We determine this by comparing the customer's DNA in that window directly against the DNA of our reference panel members in the same window. Our comparison focuses on how many members of the reference panel the customer matches, and the length of those matches.

Measuring relatedness to the reference panel using match length

Within one window, a customer's DNA is a short sequence of alleles inherited from one parent and some of their ancestors. The same is true for the corresponding DNA window in every member of the reference panel. In a single DNA window, if the customer and a reference panel member both inherited a stretch of DNA from a shared ancestor, they will carry identical matching DNA for some length of that window.

How much identical matching DNA they share depends on how many generations back in time their shared ancestor is. Each generation, recombination breaks inherited segments of DNA into shorter pieces, so DNA inherited from a recent shared ancestor is still identical over a long stretch, while DNA inherited from one further back in time is identical over only a short stretch. The length of a shared identical DNA segment in two people is therefore an indirect measure of how closely related they are in this part of the genome.

Our algorithm directly compares the DNA sequence of a customer in a specific window against multiple distinct sequences in the reference panel to determine shared matching DNA (Figure 3.2). The matching DNA segments are scored based on characteristics of the segment. Every reference sequence is labeled with the ancestral region from which it came. The collection of scored segments is used to train a classifier. The classifier outputs a probability that this segment of DNA was inherited from any of the ancestral regions. These outputs can be used to indicate, for that specific window, which ancestral region the customer's DNA is most similar to and therefore which region it is most likely inherited from.

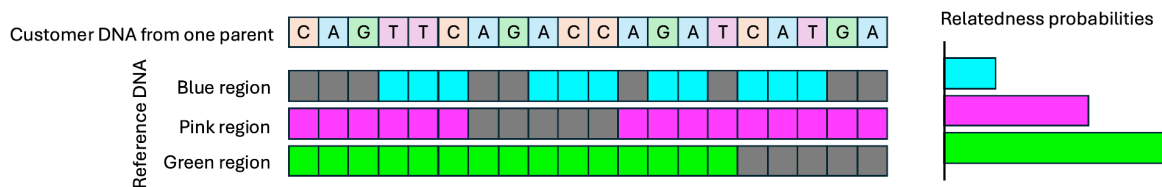


Figure 3.2: Comparing one window of a customer's DNA against the reference panel. The top row shows the DNA a customer inherited from one parent within a single window, and the rows beneath show DNA from reference panel samples from three regions, colored by region. Colored shading marks the stretch over which each reference sample matches and gray indicates places where the DNA does not match. Because DNA is broken into shorter pieces with each generation of recombination, a long matching segment indicates a more recent shared ancestor while a short one indicates a more distant relationship. The length of the matching segment measures how closely related the two individuals are in this part of the genome. Longer matching segments therefore contribute more to the resulting probabilities that a piece of DNA is inherited from that region, shown at right. The region of the long-matching reference sample has the highest probability.

Because the comparison is performed against individual reference samples, multiple samples with similar DNA sequences but different population labels remain distinguishable, with each sample contributing its own evidence. This enables the method to distinguish between closely related populations.

Tolerating genotyping error to preserve matching

Genotyping is highly accurate but not perfect, and at the scale of hundreds of thousands of positions an occasional incorrect allele is unavoidable. One or two incorrectly genotyped alleles in the middle of a long genuine shared identical DNA segment would otherwise split it into two short segments and reduce most of its value.

Our scoring therefore allows for shared identical DNA segments to include a small number of disagreeing positions (Figure 3.3). This means that a long, genuine shared segment interrupted by one error is still recognised as a long shared segment. Without this, the quality of a customer's results would depend to

an unreasonable level on the quality of their individual genotyping rather than their actual ancestral origins.

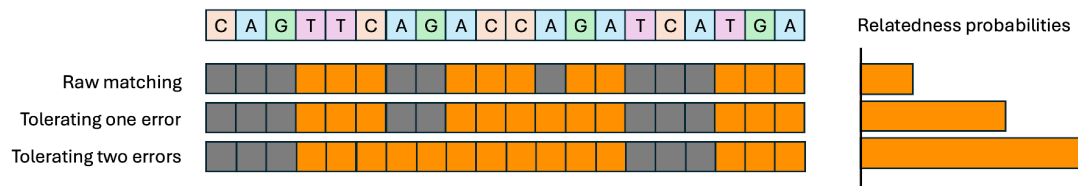


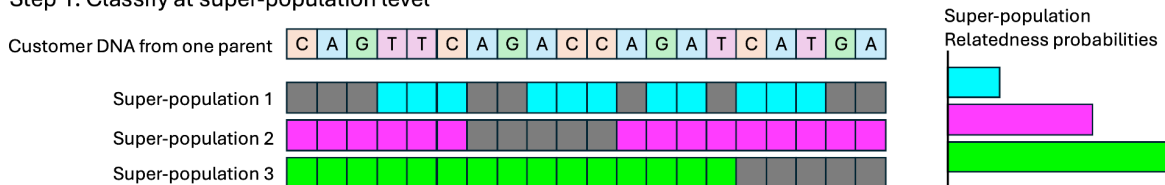
Figure 3.3: Tolerating matching errors in comparing a customer's DNA against the reference panel. Genotyping errors can split up long matching segments of DNA and reduce their value when comparing against the reference panel. A single incorrect allele in the middle of a long genuine agreement would otherwise split it into two short matching segments and discard most of its value. To manage this our scoring extends a matching DNA segment across a small number of disagreeing positions.

3.4 An Initial Super-Population Level Assignment Improves the Accuracy and Speed of the Region Assignments

Considering 302 ancestral regions as possible assignments for every DNA window adds substantial performance and computational burdens. This is especially true since many of the regions are close neighbors. For example, there are six neighboring regions in England, eight neighboring regions in Nigeria, and 15 neighboring regions in Mexico. Asking the algorithm to pick the ancestral origin of a window by considering all 302 regions at once makes these fine distinctions harder than they need to be, because each is weighed against hundreds of alternatives rather than against its genuine competitors.

The region assignment algorithm therefore works in two passes, outlined in Figure 3.4. The first pass establishes the broad picture, assigning each window to one of a small number of super-populations broadly aligning with continental groupings. The second pass gathers the windows belonging to each super-population and reconsiders them together, asking the finer question only among the regions inside that group. For example, given that we know this stretch of DNA is European, which region in Europe does it most resemble?

Step 1: Classify at super-population level



Step 2: Classify at region level



Figure 3.4: Hierarchical classification using super-population groups to constrain region assignments and improve

performance. *We first compare the customer's DNA to a reference panel of super-populations for each window. Once we have identified the most-likely super-population, we next compare the customer's DNA against more closely related regions within that super-population. The result of this step is the raw region prediction for the window.*

This is possible because the method works window by window. Since the broad origins of each window is already established, all of a customer's windows from one part of the world can be brought together and compared as a set, which distinguishes closely neighboring regions considerably better than assessing each window against every region at once.

3.5 Smoothing Region Assignments by Leveraging Information from Across the Genome

A single DNA window, considered by itself, is a limited piece of evidence about a customer's ancestral origins. In some cases, for a specific window, the probabilities are nearly identical that the DNA is inherited from one population versus another. This may be because the populations genuinely carry identical DNA in this window, or because of random chance. As a result, simply taking the highest probability region assignment in each window independently can produce a very "noisy" estimate of the customer's ancestral origins. The sequence of windows across their DNA could include frequent transitions between region assignments that are unlikely to be correct.

Patterns of how DNA is inherited can help us correct for the noisy results in our initial inference. We know that DNA is inherited in long segments and neighboring windows usually share an ancestral origin. A window whose own evidence is weak can be resolved by leveraging information from the windows around it, and a single window differing from a long run of the same population around it is more likely to be noise than a genuine change in ancestral regions.

A trained sequence model

Our region assignment model combines information across windows using an AI transformer architecture trained on data where the correct answer is known. Transformers are the model family behind recent advances in large language models and processing long sequences of text. Fundamentally, the problem we face when interpreting our windowed DNA assignments is the same as when deciphering a written sentence or paragraph: a long sequence of positions, each of them ambiguous alone, but interpretable from its broader context.

The model receives the full sequence of per-window population probabilities from our initial inference (as described in Section 3.3) and returns a "smoothed" sequence (Figure 3.5). First, every window draws on information from its surrounding neighborhood of windows. This means that evidence from a well-determined stretch of DNA can inform an ambiguous one.

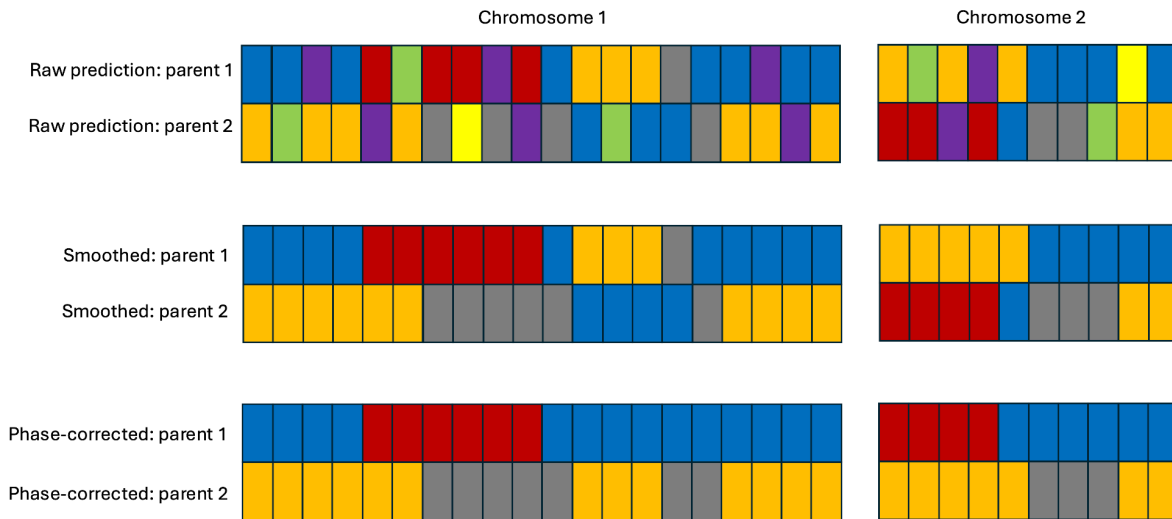


Figure 3.5: Genome-wide smoothing and phase correction improve customer overall results. Each row is a simplified illustration of a customer's DNA, where each block is a window of DNA, color coded by its assignment to one of seven regions. The top row ("Raw prediction") illustrates how raw predictions along a chromosome can be noisy because the raw predictions only leverage the information from one window. The middle row ("Smoothed") shows how we leverage information across the genome to smooth out raw prediction noise, reducing the total number of regions and producing more confident predictions. However, at this step it is possible that the parental phase may still be incorrect. The bottom row ("Phase-corrected") illustrates how we leverage region information across windows to create the final, most parsimonious, phase-corrected predictions for the genome.

Second, the statistical structure of inherited origins is learned from data. Training on cases where the truth is known lets the model discover how long segments of each population typically are, which pairs of populations are genuinely difficult to tell apart and which are not, and how much weight a window's own evidence deserves given what surrounds it. These are patterns of real human population history, and a model that learns them from data can capture regularities that would be difficult to anticipate and specify in advance.

Each parental-haplotype is smoothed separately

Recall that each window is actually composed of two sequences of DNA, each inherited from a different parent. Across windows, the DNA inherited from each parent is smoothed as its own sequence, independent of the other. This keeps the cost of the model proportional to the number of populations we report, rather than growing with the number of possible combinations of populations across the two copies. That is what makes it practical to report a set of regions as extensive as the one in this release.

In the following section we will describe how we determine which of the two sequences of DNA in each window belongs to which parent.

3.6 Correcting for Phase Error Based on Region Assignments

As Section 3.2 describes, SideView separates the DNA a customer inherited from each parent using the DNA they share with relatives. Because SideView relies on a customer's DNA matches, places in their

DNA where a customer has fewer matches will be separated with less certainty. Phasing can switch the two parental haplotypes partway along a chromosome. Left uncorrected, this appears in a customer's results as both parents' ancestral regions abruptly trading places.

The region assignment algorithm detects and corrects these switches, both before and after the smoothing process described in Section 3.5. The method we employ examines the two parental haplotypes along the genome and identifies positions where exchanging them produces a more consistent pair of ancestral region sequences, then applies those exchanges.

In sections of the customer's DNA where SideView has determined which parent a stretch of DNA came from, that information anchors the correction. As a result, this additional phase correction is orienting the sequence of region assignments to the correct parent, and not merely making the two parental sequences self-consistent. Where a customer has too few matches for SideView to resolve a stretch of the genome, the method separates the two copies and corrects switches between them using internal consistency alone.

3.7 From Window Assignments to Reported Percentages

The primary output of our region assignment algorithm is, for each window, the single most likely region of origin for the DNA inherited from each parent. These assignments, when laid out along the chromosomes, form the per-parent view of a customer's origins.

The percentages reported on the Regions results page are computed from those assignments. For each population, we total the genetic length of the windows assigned to it, divide by the total genetic length of the genome, and combine the two parental copies.

Redistributing low-confidence assignments

Despite the accuracy of our initial inference and smoothing steps, many customers still end up with some DNA windows assigned to very low-probability regions. Most of these reflect noise rather than genuine ancestral origins, and reporting all of them would obscure a customer's results rather than enrich them. We therefore report only regions that reach a minimum share of a customer's DNA, and we redistribute the remainder among the regions we do report, so that the reported percentages total 100%. However, because we do this on the results in aggregate and do not track how any specific DNA window has been redistributed, we label windows that carry these low-probability regions as "Unassigned" in the final Chromosome painter view.

This filtering does mean that a region sitting almost exactly at that minimum share can appear in one update and not in another due to chance. Even when the underlying DNA evidence is essentially unchanged, a very small movement in the estimate can carry it across the line. When a small region appears or disappears between updates, this is usually the reason, rather than a material change in our reading of a customer's DNA.

Reported ranges

Alongside each region's percentage we report a range, which represents our confidence in the exact percentage estimate for that region. Narrower ranges indicate a higher degree of confidence in our percentage estimate, while broader ranges indicate a lower degree of confidence. Section 4.5 explains how we obtain and validate those ranges.

4. Assessing Ancestral Regions Performance

While we are developing and optimizing the region assignment 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 simulations, 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 Simulations

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 to identify customers who share similar family histories (for more information see the [Genetic Communities white paper](#)). By aggregating 10s to 100s of thousands of family trees, we are able to identify patterns of admixture between populations that differ for each group. We can then simulate separate test and evaluation data sets of genotype information based on these admixture patterns, 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, ground-truth results. We measure four 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%. E.g. For someone that has 50% of all ancestors from France, if they have 75% France predicted, their overlap will be $\frac{75 \text{ observed}}{50 \text{ expected}} = 1.5$. When calculating overlap on a per-sample basis, this is calculated as the proportion predicted correctly divided by 100%.

- 2) **Recall** – the proportion of expected regions that are observed in the output. E.g., For 100 simulated people, their pedigrees include an ancestor from Norway. Only 75 of these people have Norway in their results. The recall for Norway would be $\frac{75 \text{ observed}}{100 \text{ expected}} = 75\%$.
- 3) **Precision** – the proportion of observed regions that are expected. E.g., For 100 simulated people, their results include Leinster, Ireland. Only 90 of these people have an ancestor from Leinster in their ground-truth pedigree used in the simulation. The precision for Leinster, Ireland would be $\frac{90 \text{ expected}}{100 \text{ observed}} = 90\%$.
- 4) **Weighted Precision** – precision weighted by including nearby regions that are likely very similar genetically. For example, someone with origins from the Norway region might also get assignments to Sweden, Denmark, and Iceland. Since these populations are very similar, we use weighted precision to mark these as “correct” interpretations of the genetic signal.

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 regions that are assigned at a very low percentage. Therefore, we use a cut-off of 2.5% assignment to a region to report on performance. Expected and observed values that include regions assigned below 2.5% have a much higher error and missing rate than those that include regions assigned above 2.5%.

We have run four different sets of simulations, and we have categorized these differences for the purpose of understanding model performance in different contexts.

- 1) **European-based simulations with DNA-chosen samples.** These simulations seed pedigrees with individuals from our reference panel *test* set whose DNA results strongly indicate they are from a single region (i.e., not admixed). The founders in the pedigrees have strong evidence that they are similar to other single origin people from that region. These pedigrees strictly use patterns of admixture and pedigree structure that we commonly see across European populations in the United States, the United Kingdom, Australia, and Canada.
- 2) **European-based simulations with tree-chosen samples.** These simulations use people whose trees suggest that they are single-origin, rather than relying on their DNA results. As a result, they might be similar to people living in that region today, but they might also have mixed genetics not visible in the tree, or represent other genetic variation in the region. These people are used as founders for pedigrees following patterns of admixture and pedigree structure that we commonly see across European populations in the United States, the United Kingdom, Australia, and Canada . The results from these simulations act as our lower-bound in terms of performance, while the simulations with DNA-chosen samples are considered our upper-bound, in terms of predicting where people have ancestors from in their tree.
- 3) **African American based simulations:** African-American history based simulations following a derived method similar to *Mooney et al.* (2023). We seed these simulations with *test*-set reference panel members whose DNA results strongly indicate they are from a single region (i.e., not admixed).

- 4) **Asia, Oceania, Latin America simulations:** These simulations are a composite of tree-based simulations from Asia and Oceania, combined with a theoretical based simulation of Latin American regions. These simulations are seeded with *test*-set reference panel members.

Here, we report the overall performance of the simulations.

Table 4.1: Overall performance results from a simulation of 44,880 individuals.

Simulation Set	Individuals Simulated	Median Overlap	Median Recall	Median Precision
European with DNA-chosen samples	15100	0.7621	1.000	0.8000
European with tree-chosen samples	14691	0.6851	1.000	0.6667
African American	2889	0.7248	1.000	0.7143
Asia, Oceania, Latin America	12200	0.8350	1.000	0.8000

Overall, we observe an increase in the median overlap values from our 2025 Ancestral Regions version to our 2026 version across all regions, along with consistent performance across all regions. Note that as more regions are added to the model, we would expect the overlap to go down (as there are increasingly more chances to call something to the wrong region). However, given the strong performance of the new region assignment algorithm, we see the number improve.

Here, we report performance metrics from a handful of our regions from our final evaluation:

Table 4.2: Individual region performance results for a selection of regions from the simulation of 44,880 individuals.

Region	Simulation Set	Median Overlap	Recall	Precision	Weighted Precision
Igboland	African American	0.8481	0.9866	0.7365	0.9993
Benin	African American	1.2213	1	0.9506	0.9992
Senegal	African American	0.51664	0.6770	0.7065	0.9891
Sierra Leone	African American	0.5921	0.7411	0.9130	1
Burundi & Rwanda	African American	1.0429	1	0.0165	1
Norway	European with tree-chosen samples	1.0306	0.9724	0.9635	0.9821
West Midlands	European with tree-chosen	1.8251	0.9923	0.3736	0.9955

	samples				
Sicily	European with tree-chosen samples	0.7362	0.9257	0.9756	0.9884
The Netherlands	European with tree-chosen samples	0.6994	0.9773	0.5311	0.9900
Northeastern Germany	European with tree-chosen samples	0.6627	0.9192	0.3737	0.9924
The Netherlands	European with DNA-chosen samples	0.9493	0.9915	0.8076	0.9976
Leinster, Ireland	European with DNA-chosen samples	0.9970	0.9578	0.6555	0.9868
Northern Poland	European with DNA-chosen samples	0.8477	0.9487	0.925	0.9937
Northwestern Italy	European with DNA-chosen samples	1.0257	0.95	0.8491	0.8826
Northeastern Germany	European with DNA-chosen samples	1.1614	0.9850	0.6773	0.9979
Northern Vietnam	Asia, Oceania, Latin America	0.9691	1	0.6105	0.9473
Armenians in Iran	Asia, Oceania, Latin America	1.0829	1	0.9047	1
Goa	Asia, Oceania, Latin America	0.8941	1	0.625	0.9687
Gulf of California (Mexico)	Asia, Oceania, Latin America	0.9080	0.8837	0.9665	1
Southeastern Turkey	Asia, Oceania, Latin America	0.8759	1	0.5762	1

Overall, we see very strong performance across all regions, with most having median overlap, recall and weighted precision values greater than 90%. In some regions—like *Burundi & Rawanda*, *West Midlands*, and *Northeast Germany*—we see precision values below 50%. This is expected given the total number of

regions that are possible to assign a section of DNA to. However, when accounting for assignments to neighboring regions, the weighted precision is above 90% for the majority of regions.

As mentioned previously, simulations that use people whose trees suggest that they are single-origin (tree-chosen) act as our lower-bound in terms of performance. Conversely, the simulations that identify single-origin individuals based on DNA results (DNA-chosen) are considered our upper-bound in terms of predicting where people have ancestors from in their tree. As an illustration of this point, we see that the performance for *The Netherlands* region is lower in the tree-chosen simulations than in the DNA-chosen simulations (Median Overlap: 0.6994 v. 0.9493, Precision: 0.5311 v. 0.8076). Importantly, the weighted precision in both simulations are roughly equivalent (0.9900 v. 0.9976).

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 dataset of up to 500 people to train the transformer-based smoother of our assignment algorithm, and a *test* 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 smoother *training* and *test* 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, recall, precision, and weighted precision as before. After fully tuning our model, we measured the following performance metrics (Table 4.3). We have grouped the region results at their super-population level to summarize overall trends and for readability.

Table 4.3: Results from 105,171 single-origin evaluation individuals grouped at the super-population level.

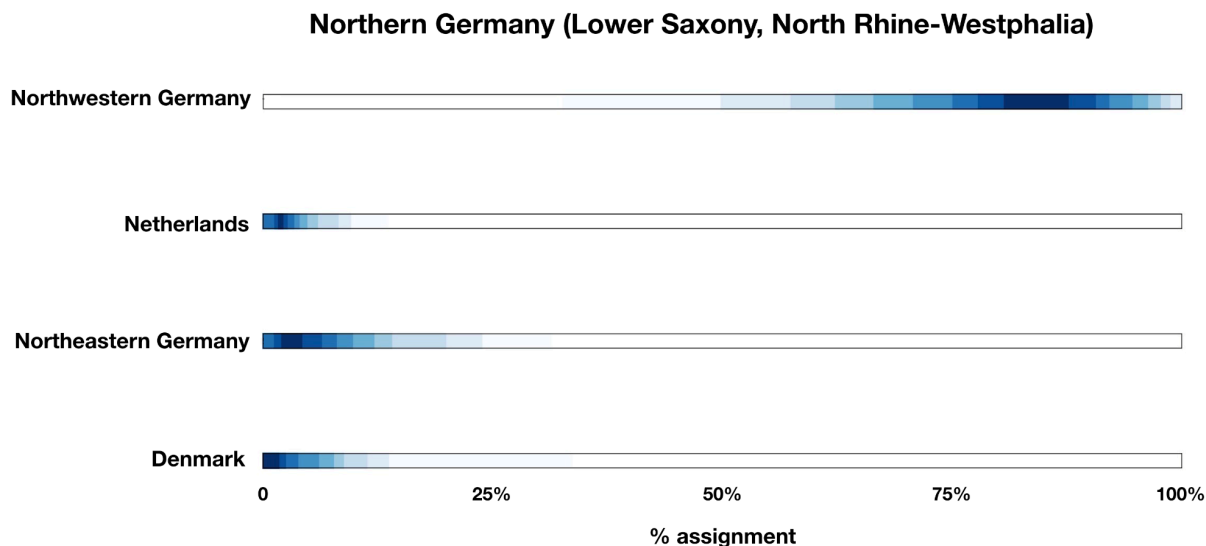
Super-population	Single-origin Individuals	Median Overlap	Median Recall	Median Precision	Median Weighted Precision
Mediterranean	24800	0.9369	1	0.5	0.9415
Africa	4264	0.8934	1	0.5	0.9657
Americas	9010	0.9833	1	1	0.9804
East Asia	9103	0.9161	1	0.5	0.9665
Europe	21634	0.9167	1	0.5	0.9044
Jewish	9330	0.9353	1	0.5	0.9727
Oceania	416	0.9713	1	0.5	0.9615
South Asia	7591	0.9603	1	0.5	0.9231
Southeast Asia	6223	0.9114	1	0.5	0.9596
West Asia	12800	0.9648	1	0.5	0.9486

Overall, we found that the majority of our new and updated regions have median overlap and recall near or above 90%. The median raw precision is consistently close to 50%. This is expected given the total number of regions that are possible to assign a section of DNA to, even when only choosing from other regions within the same super-population group. However, when accounting for assignments to neighboring regions, the weighted precision is above 90% for the majority of regions. In summary, although our models are tuned based on admixed individuals, we are pleased to report a consistently strong experience 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 the more than 500 regions of the world that we considered. For example, Figure 4.1 shows a heatmap distribution plot of the average assignments for approximately 200 customers with all four grandparents (or older) born in Northern Germany (top) and approximately 200 customers with all four grandparents born in Norway (bottom). The darker colors indicate a higher proportion of individuals with that region and that level of assignment. As you can see for both sets of individuals, while most of their assignment is to the expected corresponding regions—*Northwestern Germany* and *Norway* respectively—other regions do appear in small but appreciable amounts. These analyses help ensure that results for people from a geographic area agree with expectations.



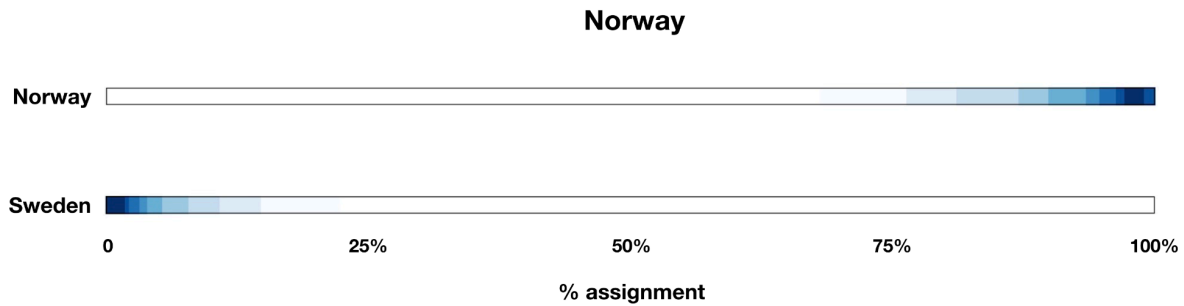


Figure 4.1 Average assignments based on grandparents' birth location. A heatmap distribution plot of the region assignments for customers with all four grandparents born in the same country or region. Northern Germany (top), and Norway (bottom). Dark blue is the middle 50th percentile, with the distribution bucketed and colored by percentile. The analysis indicates that on average individuals with roots to northern Germany will see ~80% assignment to our Northwestern Germany region and ~5% assignment to our Netherlands region. Those with roots to Norway will see ~95% assignment to our Norway region. In both cases, individuals tend to see some small percentage assignment to other neighboring regions.

4.4 Regional Polygon Construction

The process we use to create polygons for each of our 302 regions also helps to validate our model. Where possible, we use the known geographic locations of our samples to guide how we create the region polygons. Figure 4.2 shows an example of the results and geographic information used to define the polygon for our *Southeastern England* region.

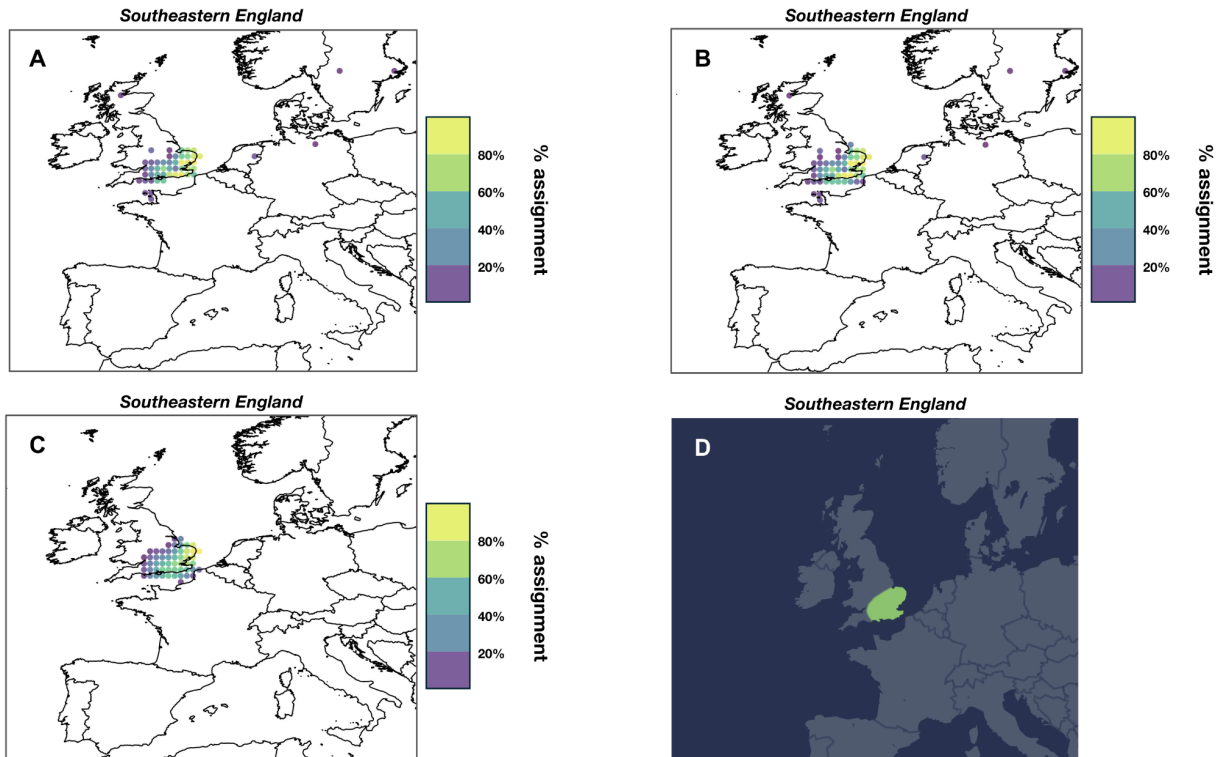


Figure 4.2: Using geographic sample locations to draw region polygons. Panel A shows the distribution of the Southeastern England region predicted for a set of samples with geographic information. Samples are assigned to grids of 0.5 degrees longitude by 0.5 degrees latitude based on the average birth location of their grandparents. The color of each grid point on the map represents the average percentage of assignment to our Southeastern England region for samples from each grid. Panel B shows the map after filling in missing grids using an imputation method. Panel C shows the information processed with further smoothing, creating the outlines representing the ancestral 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 *Southeastern England* region assigned to a combination of samples from our reference panel and *validation* set, as well as some customers with deep roots from the same country. Figure 4.2B and C show 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 southeastern England and quickly tapers off in surrounding areas. For example, the highest level of assignment, represented by a green-yellow in Figure 4.2C, is in the East and South East regions of England. The gradient continues to diminish as represented in purple, with the borders reaching towards the South West and West Midlands regions of England.

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 defined by DNA data, 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 3.7, we report a range for each ancestral region that we deliver to customers. For example, we might report to a customer a result of 40% *Southeastern England* with a range of 30-60%. This means that our most likely estimate is 40%, and that a value anywhere between 30% and 60% is also consistent with what we can determine from their DNA.

We obtain these ranges by measuring how accurate the method is, region by region. Using large sets of simulated individuals whose ancestry we constructed and therefore know exactly, we compare our estimate for each region against the known truth many times over, and we measure how far the estimate typically falls from it. That measured spread becomes the range we report.

Because the ranges come from measured accuracy rather than from an assumption about how the data behaves, they reflect our genuine confidence in assigning each region. A region whose DNA is distinctive is identified confidently and receives a narrow range. A region whose DNA closely resembles that of its neighbors, usually because the populations exchanged DNA over long periods, receives a wider one. That width is directly useful: it indicates how much weight to place on that particular value.

Ranges are reported so that they never fall below 0% or exceed 100%, since neither is a possible share of a customer's DNA. We validate the ranges against our simulation datasets to confirm that the true value falls within the reported range most of the time.

5. Future Refinement

AncestryDNA is extremely proud of the updates in this release and plans to 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. Each new release of genetic ancestral regions will represent a step forward in our ability to give our customers a complete description of their heritage 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 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. Kermamy, 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.