Ancestral Journeys White Paper: Predicting fine-scale ancestral origins from the genetic sharing patterns among millions of individuals

Last updated February 01, 2024

Ross Curtis, Ameen Eetemadi, Jenna Lang, Keith Noto, Milos Pavlovic, Chaz Reid, Alisa Sedghifar, Linhua Wang, Aaron Wolf (in alphabetical order)

Summary:

The AncestryDNA® Science team has developed a fast, sophisticated, and accurate method for assigning customers to groups who share common ancestors and historical origins from the past ~50-300 years. We identify these groups based on shared, identical, segments of DNA among our millions of customers–termed identical-by-descent (IBD) DNA segments. We quickly assign new customers to these groups with machine learning classifiers. We call this technology **Genetic Communities™**, and the feature it powers is ancestral journeys ("journeys").

Genetic Communities combines information about connections between customers from their DNA along with information about ancestral birth locations from their family trees. We simultaneously analyze up to tens of billions of DNA connections identified between several million AncestryDNA customers as a large genetic network. Patterns of connectedness in this network represent recent shared history.

By then examining the DNA and family trees of highly interconnected groups of customers, we find common ancestral origins or destinations. For example, we can find genetically related groups of customers descended from Irish immigrants to the United States fleeing the Great Famine in the 1800s.

The result is thousands of DNA based ancestral journeys that pinpoint specific locations customers' ancestors may have lived and moved over time. We combine these assignments with historical insights about the people who lived in that region and the forces that shaped their lives.

1. Introduction

AncestryDNA® offers several genetic analyses to help customers discover, preserve, and share their family history. Some of the features offered to date are based exclusively on genetic information. These include genetic inference of ancestral regions (described in *Ancestral Regions[White Paper](#)*) and an identity-by-descent (IBD) or DNA matching analysis (*[Matching White Paper](#)*). Each of these features provides complementary information to a customer: (1) the "Ancestral Regions" provides a distant picture of the customer's genetic origins, perhaps hundreds or thousands of years ago; (2) the "DNA Matches" feature provides a customer with a list of fellow AncestryDNA test takers that are relatives with whom she or he shares a common ancestor within the last 10 generations.

Here, we augment these DNA-based insights even further with a technology we call **Genetic Communities™** (Figure 1.1). Instead of considering the IBD connection between each pair of customers in isolation, we simultaneously analyze up to tens of billions of connections identified among several million AncestryDNA customers as a large genetic network (described below in section 3). Intuitively, because the estimated IBD connections between individuals are likely due to recent shared ancestry (within the past 10 generations), broader patterns in this large network likely represent recent shared history. The result is that we can identify clusters of living individuals that share large amounts of DNA due to specific, recent shared history.

For example, we identify groups of customers that are likely descended from immigrants participating in a particular wave of migration (e.g., Irish fleeing the Great Famine), or customers that are descended from ancestral populations that have remained in the same geographic location for many generations (e.g., the early European settlers of the Appalachian mountains). Following the identification of these clusters of individuals in the entire network, we can then assign any AncestryDNA customer to one or more of these clusters based on their IBD with other AncestryDNA members. These assignments can provide a customer with insight into their recent ancestral history, in some cases traceable back to a historical event.

In the following sections, we describe the scientific principles behind the genetic network (Sections 2 and 3), how we identify clusters within it (Sections 4 and 6), our use of DNA and pedigree data to annotate these clusters (Section 5), our method for assigning customer samples to these clusters (Section 7), and how we infer the parent-of-origin for individuals' assignments (Section 8).
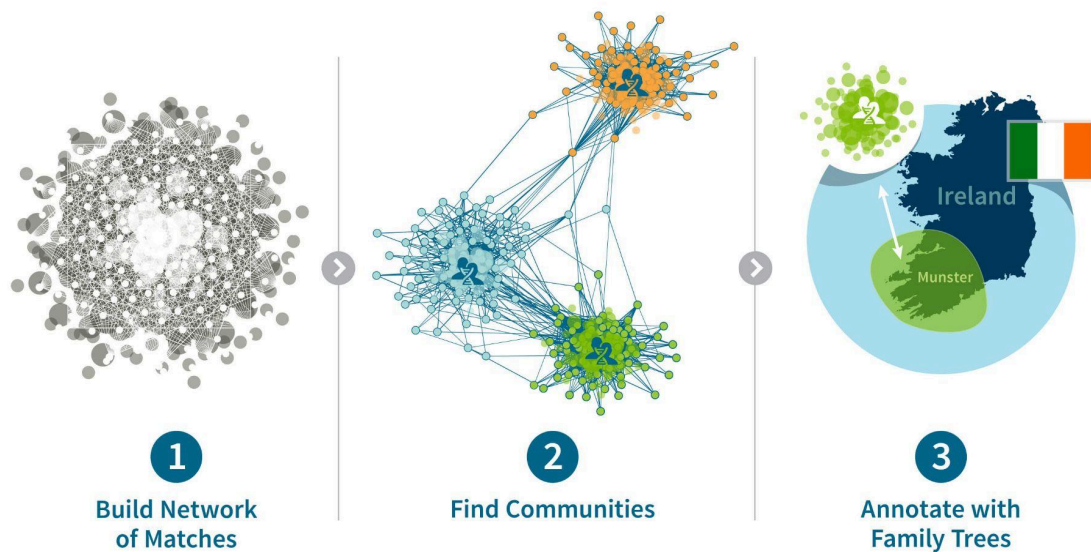
Figure 1.1: An overview of how we identify ancestral journeys using our DNA database.

2. Population Genetics Motivation for Ancestral Journeys

In this section, we will discuss some basic population genetics concepts that motivate the development of Genetic Communities™, and conclude with an example of one such ancestral journey.

First, we introduce some terminology. An **IBD network** is a representation of the genetic connections among a collection of AncestryDNA samples. The nodes in the network are the samples, and the edges between nodes are the IBD connections between samples, i.e. the amount of matching DNA shared by two samples. We describe the IBD network concept in detail in Section 3.

A **community** can be thought of as a part of the larger network that has a high degree of connectivity. Specifically, nodes in a community will have higher rates of IBD (and longer segments of IBD) with other nodes *inside* the community than they do with other nodes *outside* the community. See Section 4 for a broader discussion of communities. Ultimately, using genetic and historical information, we infer the shared ancestral journey that has produced this genetic community within the IBD network.

To understand why we expect to find community structures in an IBD network, we examine a few basic population genetics principles.
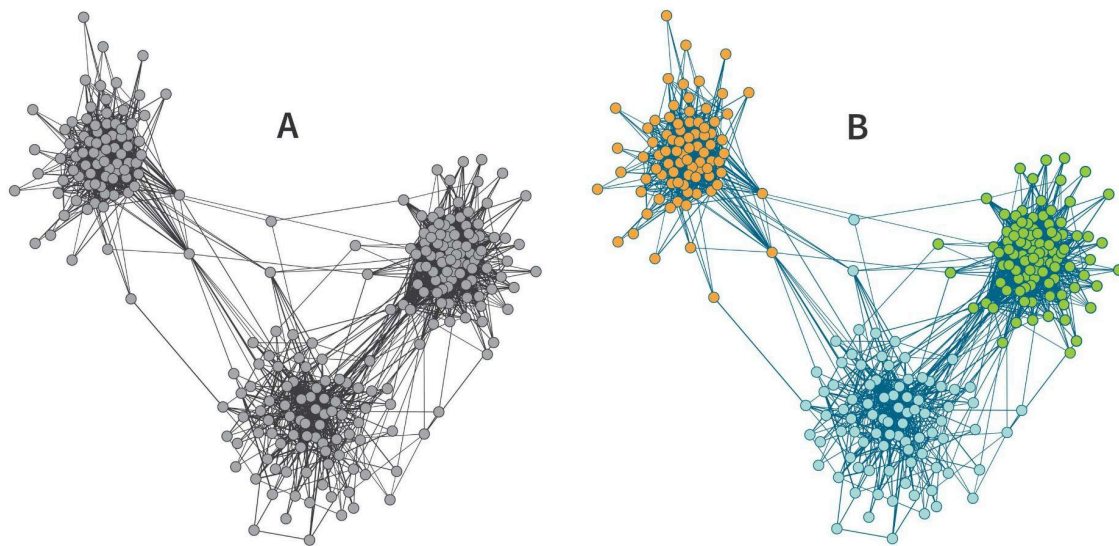
Figure 2.1: An illustrated example of an IBD network and community clusters. (A) In this IBD network, the circles are nodes and represent AncestryDNA samples. The lines between samples represent IBD connections, weighted by the amount of DNA shared. (B) There are three communities discovered in this network, colored orange, green, and blue. While samples in these communities have some connections to nodes in other communities, they have a higher rate of connectivity to samples within their own community.

2.1 Genetic Populations

We begin with a discussion around the concept of a genetic **population**. Numerous definitions of what constitutes a population exist in the genetics literature. For clarity, we define a population as a group of people who generally live in close proximity and produce children with one another for multiple generations. This definition is intentionally vague with regards to size and scale. A population can be a large, loosely connected group, such as all *Europeans*, or it can be a smaller, more closely connected group, such as the *Irish*. While vague in terms of scale, our definition of a population is specific with regards to time and place. For example, one population might include the ancestors of Europe that lived ten thousand years ago, while another population might include people living in Connecticut 200 years ago.

Each population has a different **degree of genetic isolation**. When a population has a high degree of genetic isolation, that implies that population members rarely if ever have children with individuals outside the population. On the other hand, a population with a low degree of isolation has high levels of movement and admixture with surrounding populations. Over time, isolated populations develop distinguishable patterns of genetic variation.

New populations can be created in numerous ways. For example, a small subset of individuals from a historical population may migrate to a new location and create a new population that no longer produces offspring with the source population. It is also possible for this new population to separate from the historical population without leaving the source location. Another possibility is for multiple source populations to come together and **admix**, producing offspring with a mixture of genetic material from formerly separated populations. In all these examples, the unifying feature is that a **barrier to gene flow** between groups leads to the development of distinguishing patterns of genetic variation.

There are many forces including geography, war, religion, culture, politics, and economics that may influence whether populations admix or remain separate from each other. These forces have a significant impact on how DNA is shared among individuals over time and space. In our work, we demonstrate that it is possible to observe these historical forces, which influenced our ancestors, by examining the DNA of individuals living today.

2.2 An Illustrative Example

Let's consider a simple example that demonstrates how admixture in a population can lead to community structure in an IBD network.

In Figure 2.2, we represent ten unrelated individuals from ten different populations. Note that the ten founding individuals do not share any IBD segments of DNA since they are from different populations (denoted by the different colored bars). In this example, these ten individuals randomly mate and each couple (there are a total of five couples) has two children, creating a second generation of ten individuals. In this second generation, some individuals are now loosely connected by IBD at the close family level (i.e., they share color patterns).

We keep repeating this experiment for two more generations; ten unrelated individuals in the second generation randomly choose mates and each couple has two children, creating a third generation of ten individuals. Finally, ten individuals in the third generation randomly mate and each couple has only one child, creating a fourth generation of five individuals.

Interestingly, after these three generations of random mating, all five progeny in the fourth generation have at least one part of their ancestral makeup that is shared with each of their other four cousins. These five individuals also have a higher rate of IBD with individuals in this new population than they would with people from the original ten ancestral populations. In an IBD network made up of individuals from many other populations as well, this particular population would likely form a community.
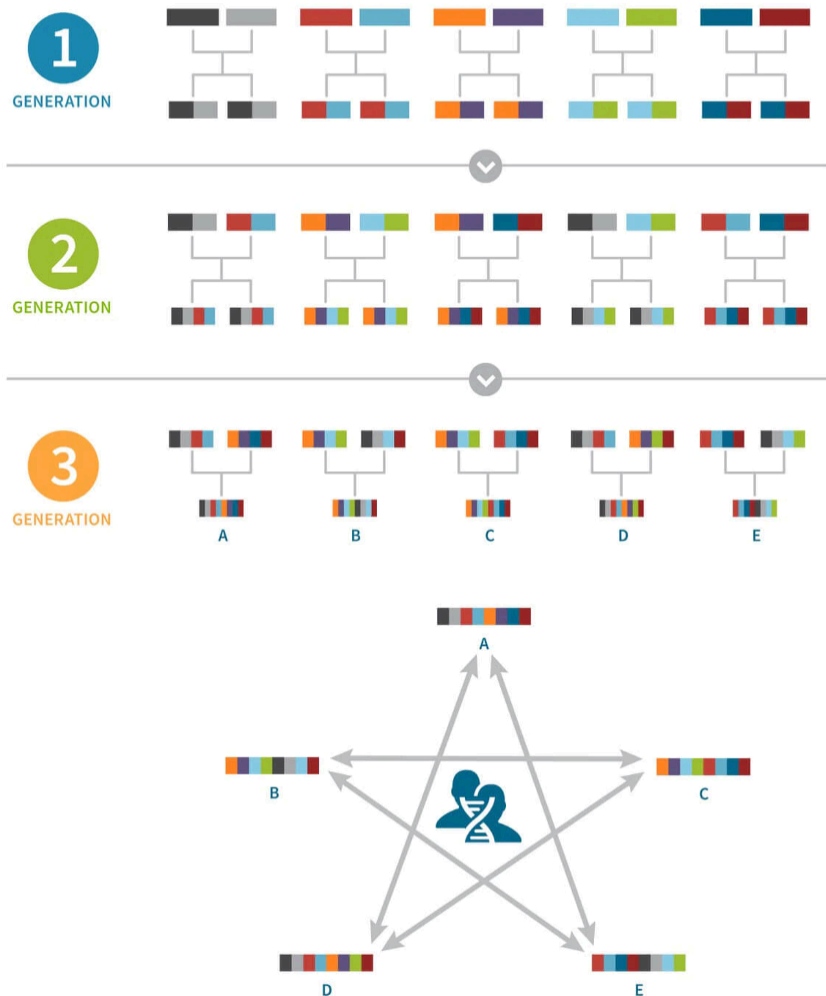
Figure 2.2: An illustrative example of the structures in an IBD network. Each individual is represented by a single bar, which is colored according to the individual's ancestral makeup. Note that in generations one and two, there are five couples that each have two children. In generation three, each couple only has one child. Due to the shared history of these five individuals, they would all have an IBD relationship with every other individual in the population, denoted by the shared color patterns. This creates a completely connected network with relationships depicted by double arrows.

While this example is exaggerated in its simplicity, it helps to illustrate the intuition behind populations and how genetic isolation and admixture can create structure in a large IBD network. Of course, real populations typically have hundreds or thousands of founders and are generally not completely isolated. The degree of admixture and migration in a population will affect the strength of its structure in the IBD network.

Also, while the IBD network in our example is completely connected in the fourth generation, in

large populations we will rarely find a completely connected network. Rather, it is the presence of higher rates of IBD among individuals in the same population due to the intermarriage of hundreds or thousands of families over the course of many generations that creates a modular structure in the network. When this occurs, individuals have more IBD connections to other individuals in the same community (or population) than they do to individuals from other communities in the larger network.

2.3 Example of a Recent Population in West Virginia

We conclude Section 2 by discussing the creation of a population that settled in western Virginia during the 18th century and use it to provide the intuition that is the basis of our Genetic Communities™ technology and ancestral journeys feature (Figure 2.3).
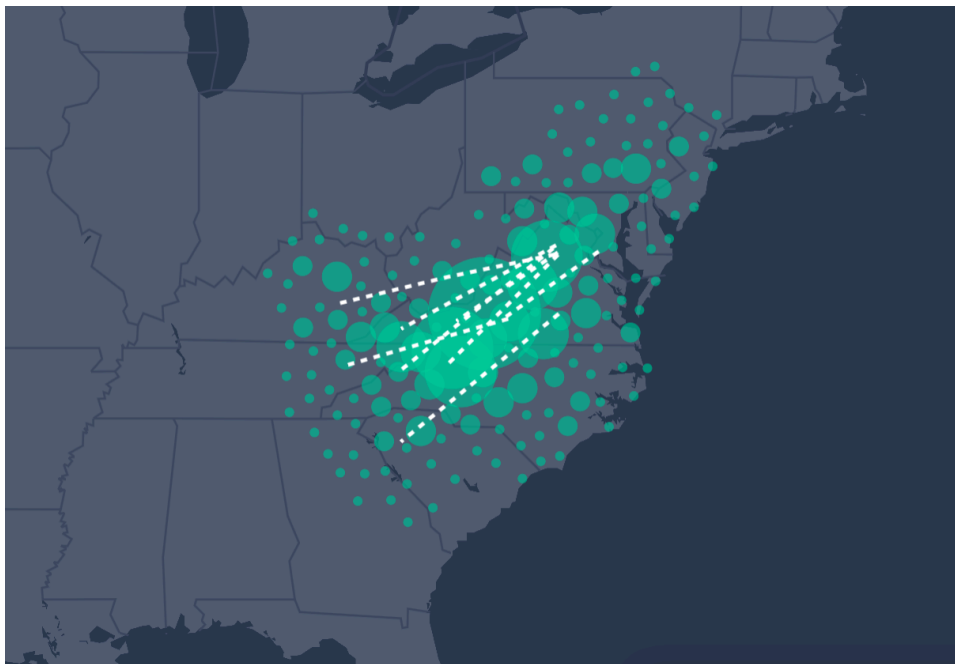

Figure 2.3: An ancestral journey forms in western Virginia in the 1800s when ancestors of the group move to the same region. This journey represents a case where founders from different regions converged on the same area in Appalachia around the same time. The green circles on the map refer to the birth locations for the population founders, spread across the mid-Atlantic and Northeast. The white lines indicate shared patterns of migration from coastal Virginia to settlements in the Appalachian foothills in the early 1800s. Data are projected onto maps from OpenStreetMap (openstreetmap.org/copyright)

2.3.1 The history of West Virginia settlement
European-American settlement of western Virginia and West Virginia first began in the 1730s, when the Virginia colony promoted settlement of its western mountains to provide a buffer between its established towns and native peoples (Rice 1993). Between 1750 and 1780, the

founding population in this region grew. This was a period of peace, prosperity, and aggressive settlement in the Shenandoah Valley following the end of King George's War in 1748 and a treaty with native peoples in 1752. While the British forbade further settlement with the Proclamation of 1763, Americans pushed into the territory following the Revolutionary War. The construction of roads between 1818 and 1846 promoted further settlement and the isolation of rural areas (Rice 1993). Thus, until the mid-19th century, this region was largely a rural, growing population with settlers who hailed from British, German, or Scotch-Irish backgrounds.

Between 1850 and 1890, this region saw a period of industrialization and a corresponding population boom due to the beginning of the coal industry in West Virginia and the development of the C&O Railroad and coal towns that emerged along its route. For example, Kanawha County experienced a population growth of 700% between 1890 and 1910 (Laidley 1911 [310]). The dawn of the 20th century saw a shift in the pattern with a massive out-migration following World War I to the industrial cities of the Midwest and West.

2.3.2 Discussion about West Virginia settlement

In this example, we see a new population created in the latter half of the 1700s, consisting of founding individuals from Scotch-Irish, German, and British heritage.

While descendants of this population will surely carry DNA indicating a link to their distant Scotch-Irish, German, and British origins, mating between founders of this new population, and subsequently their descendants over many generations, has resulted in the formation of a new population with patterns of genetic variation that are related to, but distinct from, their historical source populations. The descendants of this new population are people who share large amounts of genetic material with many other descendants in this population. Families intermarried throughout the 1800s until people began to leave after World War I. However, even for the descendants of families that left West Virginia some time ago, the genetic signature persists in the form of long IBD segments shared among descendants regardless of their more recent familial history. Thus, we expect to discover this group of descendants in our AncestryDNA database using the IBD connections between these individuals. In the following sections, we will show how we discover this and other descendant populations in an IBD network.

It is important to note that the examples in this section are intended to illustrate general principles motivating our approach to use community detection to discover communities in a large IBD network. These two examples do not represent the unique history of all populations around the world. Each community that we discover has its own unique history and degree of genetic isolation and migration. That being said, some of the principles we have discussed will apply in many populations. Furthermore, we can use genetic and historical information to infer the ancestral journey of these communities.

3. Constructing an IBD Network from IBD Connections

In this and the subsequent sections, we introduce the methods we use to discover and annotate ancestral journeys.

We begin with the collection of all pairwise IBD connections identified between AncestryDNA customers. A pair of customers is said to have an IBD connection if they share one or more long segments of identical DNA. The most likely explanation for a long segment of identical DNA present in two individuals is that it has been inherited by both individuals from a single common ancestor and thus indicates 'identity-by-descent' in the two descendants.

Let's use Customer A in Figure 3.1 as an example. We have identified, by comparing his DNA with all other customers in our database, seven other customers who have an IBD connection to him (B, C, D, E, F, G, and H). (See the *Matching White Paper* for more details on how we identify IBD connections.) The genetic connections in this small example can be summarized visually by drawing edges between the pairs of people we have identified as connected based on their DNA (Figure 3.1). In this particular example, A, B, and C are siblings, so all three are connected by edges.
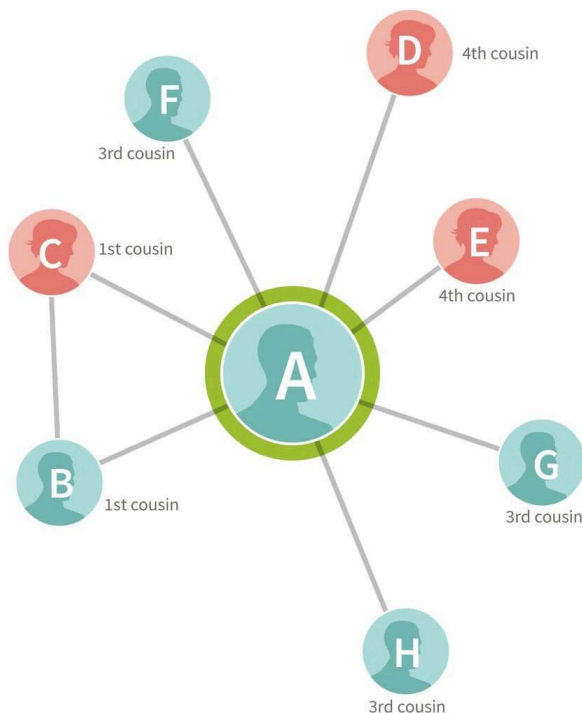

Figure 3.1: An example IBD network of genetically related AncestryDNA customers. In this figure, DNA matches between customers A, B, C, D, E, F, G, and H are shown as lines. Note that in this case, customers B and C match A and also share DNA with each other.

Next, we expand on this example by including the IBD connections found for each of A's seven connections (Figure 3.2). The size of the network expands rapidly as we add more people by following the genetic connections between them. In some cases, these new samples are also related to each other, and in other cases these new samples are connected to others that are already included in the network (the green circles). In both cases, we draw edges reflecting the identified IBD connections between these individuals.
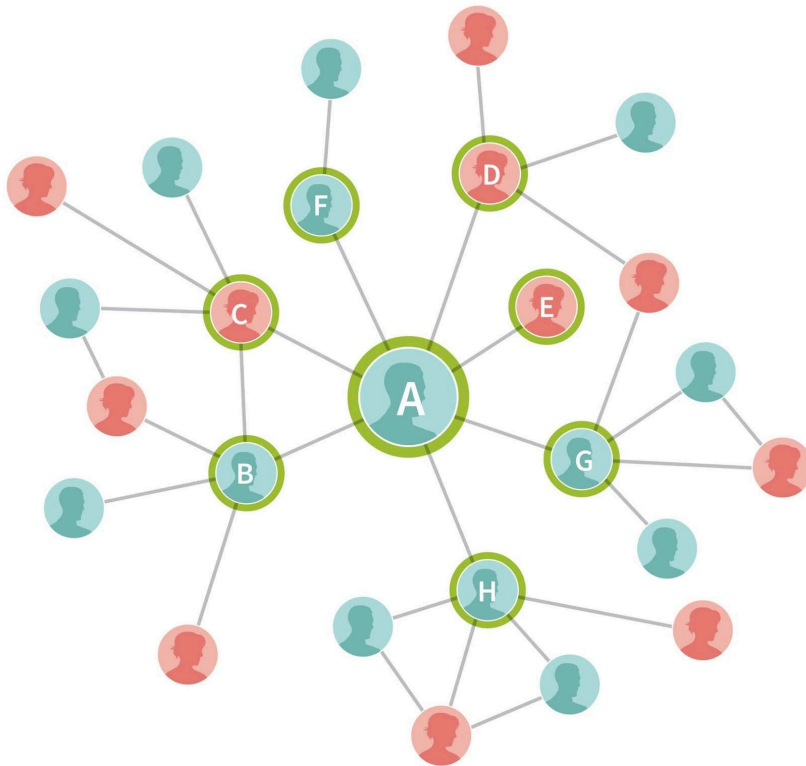
Figure 3.2: **An expanded network of genetically related AncestryDNA customers.**This figure highlights how quickly a network expands and shows that in some cases these new samples are related to each other.

Extending this logic further, we form an IBD network from the IBD connections detected among the millions of individuals that have taken the AncestryDNA test. Clearly, visualizing this network in a single figure, like we have done above, would be difficult. To illustrate what one small part of this network might look like, we show the IBD connections detected between a set of 75 selected AncestryDNA samples (Figure 3.3). This is an example of a particularly well-connected group of samples in the AncestryDNA IBD network, yet there are still pairs of people among this group for whom we did not find an IBD connection.
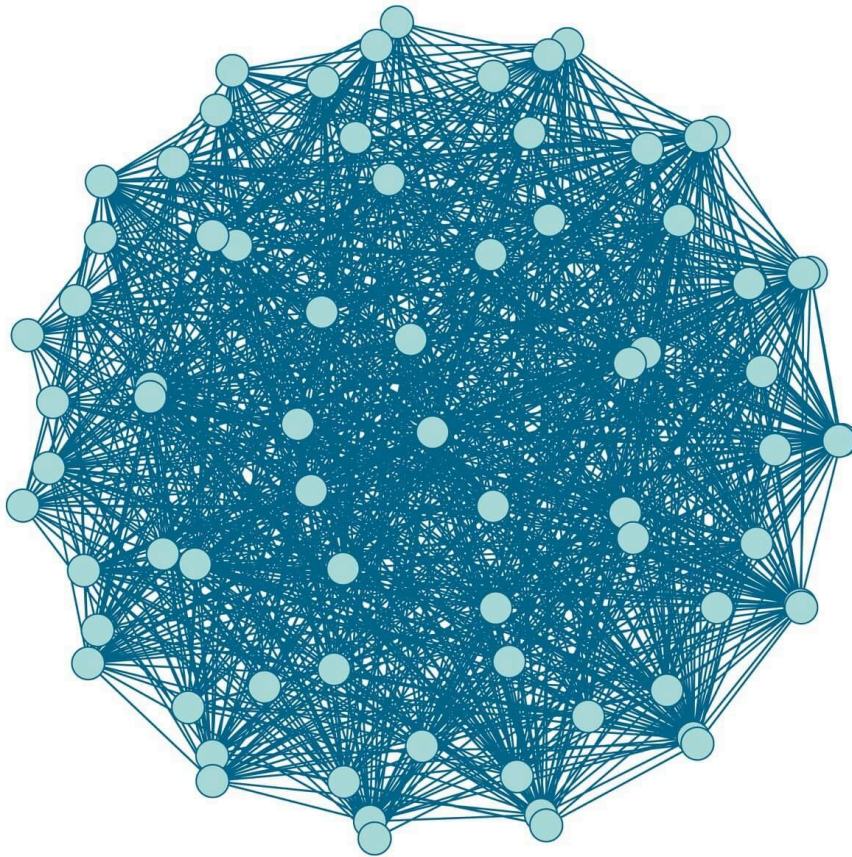
Figure 3.3: IBD connections between 75 customers selected from the AncestryDNA IBD network.

4. Network Clustering by Community Detection

Given an IBD network, we can subdivide the network into densely-connected communities using the Louvain Method–a popular community detection algorithm. Community detection algorithms are network clustering methods that identify "strongly connected" subsets of a network (Blondel et al. 2008, Csardi et al. 2008). In the case of our IBD network, these communities represent groups of individuals that are more related to one another than they are to others in the network.

Going back to our visual representation of the IBD network of 75 samples in Figure 3.3, there is no obvious pattern in this IBD network. In Figure 4.1, we present the same network with the nodes rearranged to highlight the structure in the network. In particular, the 75 individuals have been subdivided into three communities, which we have labeled as Community A (24 individuals, shown as green circles), Community B (30 individuals, orange) and Community C (21 individuals, blue). Note that these communities were not detected by visual inspection, but

rather by running a detection algorithm on the IBD network of 75 individuals, which assigns each individual to one community.
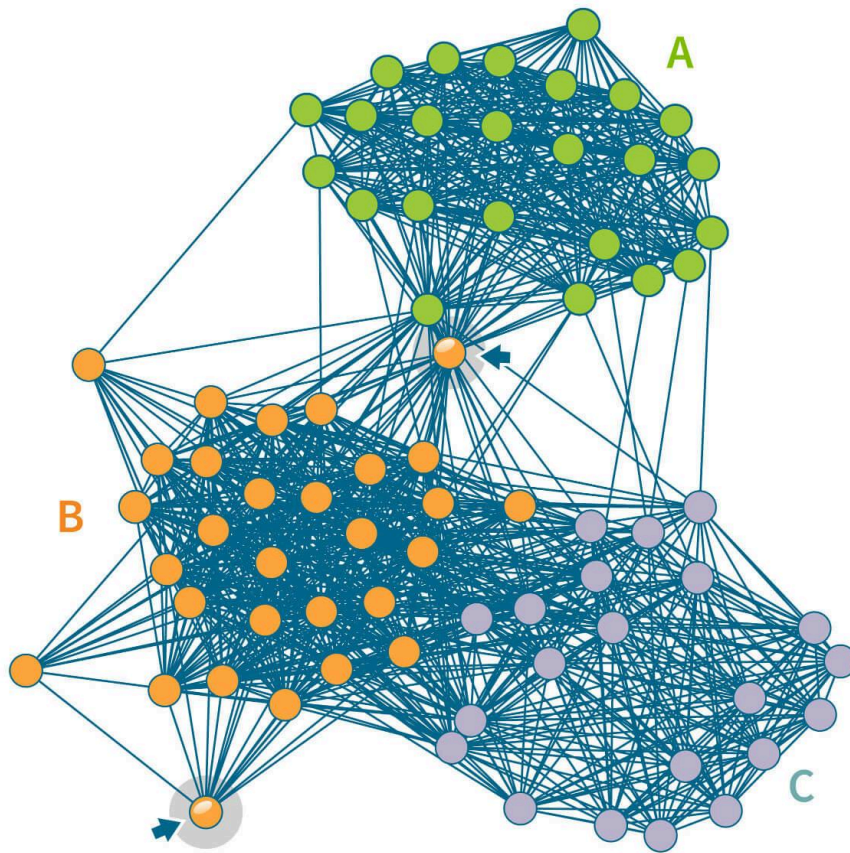

Figure 4.1: The individuals depicted in Figure 3.3 are arranged into three communities, highlighting the network structure. These communities are labeled as community A (24 individuals, shown as green circles), community B (30 individuals, orange), and community C (21 individuals, blue). The arrows highlight two individuals with different types of IBD connections: (1) the individual at the bottom of the figure has connections that are contained within a single community, and (2) the individual in the middle of the figure has connections spread across multiple communities.

At a basic level, the community detection algorithm is subdividing the network into subsets that are more densely connected than the original full network. We can measure how connected a network is using a measure called **network density** (Table 4.1). The density is the number of edges present in the network divided by the number of edges possible in the network. In the case of the IBD network, the network is maximally connected if there is an edge between every pair of individuals. Following community detection in our example above, pairs of individuals **within** the same community are more densely connected to each other than pairs of individuals **between** communities. For example, 185 edges are contained in Community C, for a density of: 185 × 2 / (20 × 21) = 88%. Whereas only 163 edges join members in Communities B and C, for

a density of: 163 / (21 × 30) = 26%.

	Cluster A (n=24)	Cluster B (n=30)	Cluster C (n=21)
Cluster A	276 (100%)	43 (6%)	9 (2%)
Cluster B		373 (86%)	163 (26%)
Cluster C			185 (88%)

Table 4.1: Number of edges within and between communities in the example IBD network.

Subdividing this network into three communities illustrates another important concept to consider when investigating patterns of IBD connections across many individuals: some individuals have most or all their IBD connections contained within one of the groups, whereas other individuals have IBD connections that are spread across multiple groups. An example of the former is a node in the bottom-left corner of Figure 4.1. The edges emanating from this node all connect to other nodes within the same community (Community B). By contrast, in the middle of the figure, we highlight an individual that is assigned to Community B even though this individual has IBD connections with many members of both communities A and B, as well as a couple from Community C. Therefore, the "degree" or "strength" of membership into a particular group is greater for some individuals than others.

By applying fast network community detection algorithms to our AncestryDNA IBD network, we are able to detect population structure within the network. In Section 6, we will discuss how we recursively run community detection to discover fine-scale population structure.

5. Turning Communities into Ancestral Journeys by Interpreting their Historical and Geographical Characteristics

Communities are discovered solely by using the IBD connections between individuals. As described in Section 2, we expect these connected communities to each represent a group of descendants of a particular ancestral population and journey. But how can we identify the historical population responsible for a particular set of connections and turn a genetic community into an ancestral journey to present to customers?

For this we rely on both genetic data and information present in the pedigrees of community descendants. In particular, since these connections reflect recent common ancestry, we can look for common features that are shared by individuals in each community to correlate the genetic patterns to recent history. These common features help identify a common time, location, or source population from which descendants have ancestry. For example, the people

in one community might be the descendants of Irish immigrants who came to the United States during the Great Famine in the 19th century. This can be represented as a compelling ancestral journey to customers.

For this analysis, we rely on two sets of data: (1) genetic admixture proportions from over 80 global populations (see our *Ancestral Regions White Paper*), and (2) pedigrees curated by the users who have taken the AncestryDNA test. The scale and diversity of these data allow us to infer detailed historical and geographic portraits of the communities detected in the IBD network.

Our ability to identify the ancestral journey shared by a particular community is strongly dependent on the data available. For example, if no community members have created pedigrees, we have limited ability to identify a source location for the journey. Also, an individual can only be linked to a particular journey if they share significant amounts of genetic material with others descending from the journey. Without a genetic connection, we are not able to link individuals to a particular journey. However, the continued growth of the AncestryDNA database positively impacts both of these limitations.

5.1. Genetic Admixture Analysis

The first feature we look at for each ancestral journey are the ancestral population proportions estimated from the DNA. These annotations can be used to estimate which ancestral populations are overrepresented or underrepresented among individuals from a given journey. In some cases, journeys with highly overrepresented ancestral populations can be related to known populations. For example, journeys corresponding to relatively recent US immigrant groups such as Finnish, Jewish, and Irish people can be identified from the ancestral region-based annotations. On the other hand, journeys corresponding to groups from New York State, Pennsylvania, and Ohio will have similar, non-distinguishing genetic profiles.

Figure 5.1 looks at the ancestral region profile of the members of a specific journey that we discover in the IBD network. The average result of these individuals is primarily from Ireland, suggesting that these individuals have a shared Irish ancestry.
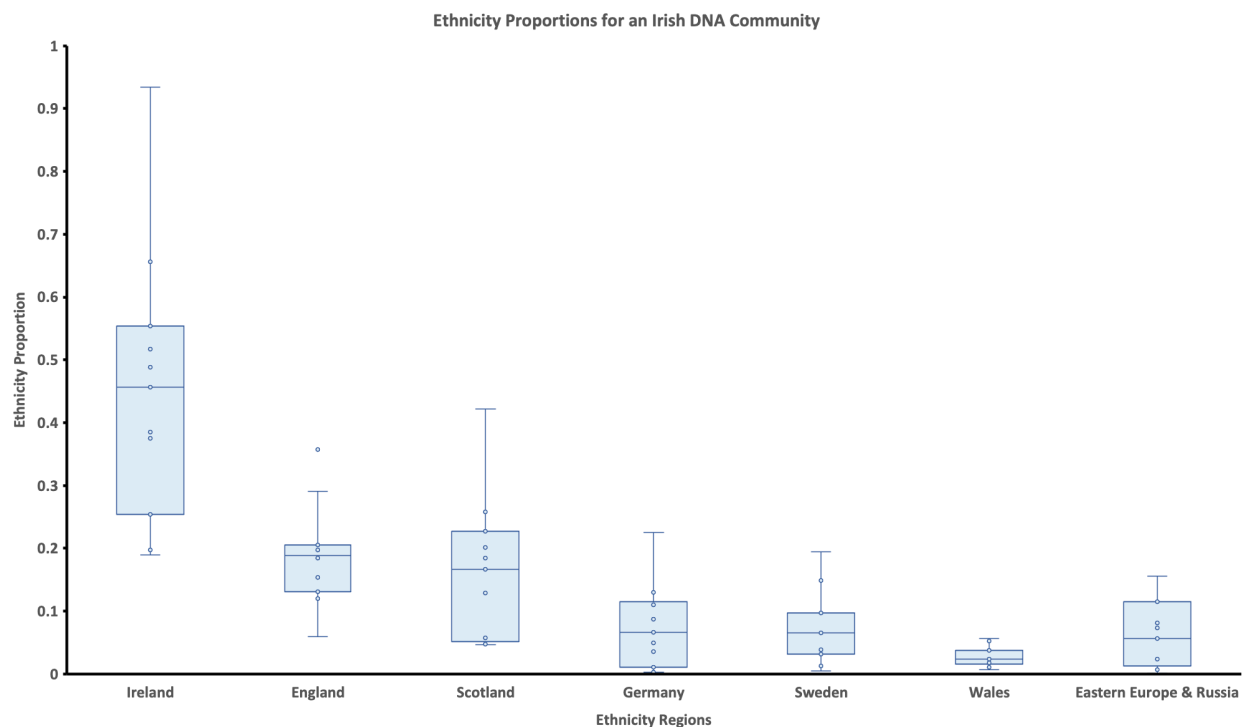
Ethnicity Proportions for an Irish DNA Community

Figure 5.1: The average ancestral region proportions for an Irish ancestral journey. This boxplot shows the ancestral region proportions for members of a specific journey. The median results of individuals in this journey is close to 50% Irish, suggesting this group has shared Irish origins.

5.2. Enriched Surnames

Next, we consider the surnames of the ancestors of journey members using aggregated pedigree data. To summarize ancestral surnames for a given journey, we collect all the surnames of recent ancestors that are associated with the individuals who are assigned to the journey. To highlight surnames that are more likely to be characteristic of the journey, and therefore more likely to yield informative clues about the historical or demographic significance of the journey, we quantify the statistical evidence (i.e., *p*-value) that each surname is over-represented in a given journey compared against the background surname distribution over all individuals in the full AncestryDNA database. Then, we rank the surnames according to the statistical evidence, and select the 20 most highly ranked surnames as the surnames that are characteristic to the given journey. For example, the highly ranked surnames from the surname annotations associated with individuals assigned to an Irish journey include "McCarthy", "Sullivan", "Murphy", "O'Brien", and "O'Connor" (Figure 5.2)
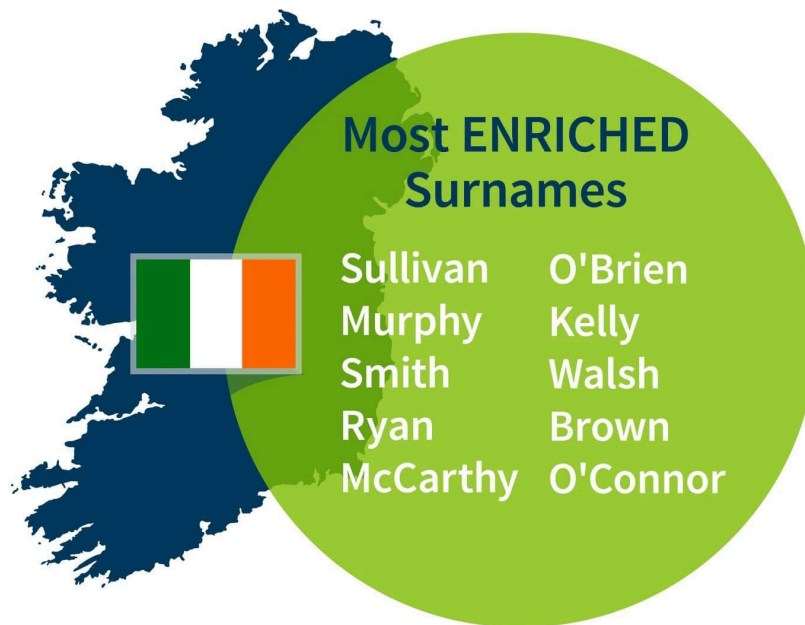
Figure 5.2: The enriched surnames for an Irish ancestral journey.

5.3. Enriched Birth Locations

Another type of annotation that we use to characterize journeys is the birth locations of the ancestors associated with individuals assigned to a journey. These locations provide useful geographic clues that often can connect a journey to historical populations. For this analysis, we compile statistics of birth locations of the ancestors specific to each journey throughout time and summarize the birth location data so that it may be visualized geographically.

This analysis is accomplished by converting each birth location, within a specified range of generations, to the nearest coordinate on a two-dimensional (2-D) grid. For each grid point in the 2-D grid, we compute an odds ratio (OR). This OR is defined as the odds that a given grid point of the 2-D grid is associated with the journey members **divided by** the odds that the same grid point is associated with users who are not members of the journey. Using this OR measure, we generate a map that visually depicts grid points in which the largest odds ratios are indicated visually by labels or distinguishable colors. In this way, the highlighted graphical map locations correspond to geographic locations that are disproportionately enriched in a given journey.

For example, Figure 5.3 shows the enriched birth locations of ancestors born between 1850 and 1910 associated with a journey with origins in Ireland. This map shows that birth locations with high OR (therefore more enriched) are more highly concentrated in the southern part of Ireland (Munster).
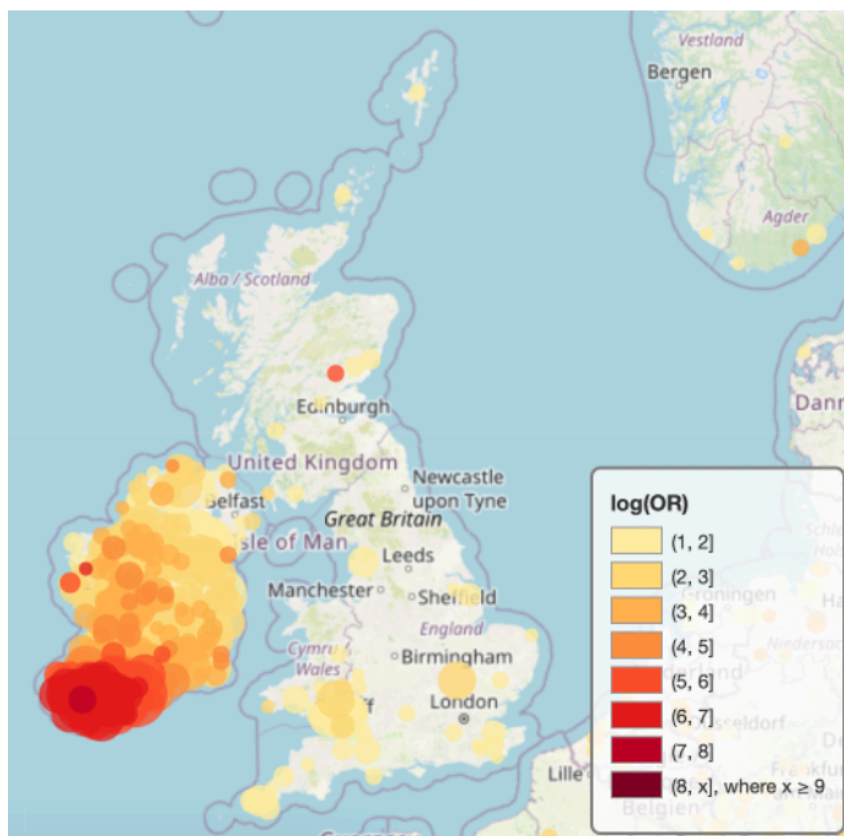
Figure 5.3: The ancestral birth location enrichment map for a journey originating in Munster, Ireland. Birth locations with a high odds ratio are colored in red (most significantly enriched) and yellow (less significantly enriched). The enriched locations are primarily in the Munster region of southern Ireland. Data are projected onto maps from OpenStreetMap (openstreetmap.org/copyright)

In addition to looking at the odds ratio, we also consider the proportion of the samples in a journey that have ancestral birth locations in the region identified for the journey. To do this, we first use the birth location enrichment plots to construct polygons around significant locations specific to each journey (these polygons are also used in the product experience).

Based on these specific locations, we can determine, for each individual assigned to the journey, which ancestors were born in this area. For example, in Figure 5.4, we show the average number of ancestors born in Munster, Ireland, by generation. When we look at the family trees of *Munster Irish* ancestral journey members, we find that around four to five generations ago (approximately 150 years ago), they have an average of one ancestor born in Munster. For individuals not assigned to this journey, exceptionally few people had an ancestor at any time born in this region. These analyses support our interpretation of this journey as the descendants of people who lived in Munster, Ireland.

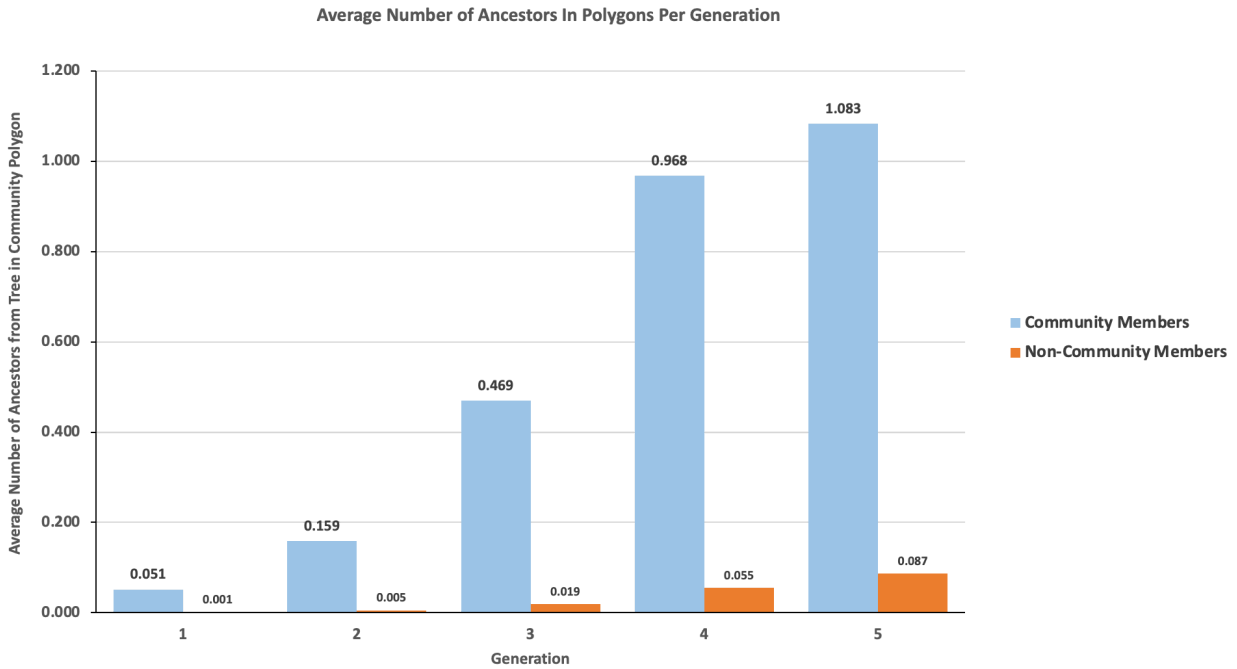**Average Number of Ancestors In Polygons Per Generation**

Figure 5.4: Comparing the number of ancestors born in Munster, Ireland, for individuals assigned to an Irish ancestral journey and those not assigned to the journey. The bars represent the average number of ancestors born in Munster, Ireland, by generation. The data are collected from family trees. Blue bars represent the numbers for people assigned to the *Munster Irish* journey while the orange bars represent the number for people not assigned to the journey.

5.4. Migration Patterns

Finally, we also study the migration patterns of the ancestors of journey members through time, as observed from the aggregated pedigree data. We examine how the ancestors of people in this journey moved from one location to another by looking at the birth locations of parents and children for each generation in each pedigree. Thus, we define a migration path as a path from the birth location of a parent to the birth location of a child.

By looking at changes in these migration paths, we often gain further insight about the population dynamics of the ancestors of the people in this journey, and how those dynamics have changed through time.

For example, when we look at the Irish journey from Munster, Ireland, we see a very high frequency of migration paths from Munster to the United States from 1825 until 1900 (Figure 5.6). This time frame corresponds to the migration of 6 million Irish to the United States in the 19th century, which peaked in 1852 during the Irish Famine (Fitzgerald and Lambkin 2008 [8,181]).

Figure 5.6: A map of migration from Munster, Ireland, to the United States circa 1875. We used ancestral birth location information from family trees to map the patterns of migration over time for members of the *Munster Irish* ancestral journey. Data are projected onto maps from OpenStreetMap (openstreetmap.org/copyright)

5.5 Ancestral Journey Interpretation

Based on these four pieces of information–ancestral regions, surnames, birth locations and migration paths–we are often able to infer some of the historical context leading to the strong genetic connections between individuals in the same journey. These interpretations are used to guide the names of the journey in the user experience, as well as associated historical insights.

6. Recursively Discovering Fine-scale Subnetworks

With millions of samples in the AncestryDNA database, recursive application of community detection identifies fine-scale structure in the IBD network. When we first applied community detection on the IBD network of samples in the AncestryDNA database, we initially identified only a handful of communities in the network, generally representing either subtle gene flow barriers affecting hundreds of thousands of samples or stronger gene flow barriers separating much smaller subsets of the IBD network.

A key discovery of this work was that it is possible to uncover smaller, higher-resolution communities through the recursive application of the community detection algorithm. Since each observed community is itself a network of IBD connections on which we can apply the same community detection algorithm to discover sub-communities, we performed community detection recursively. Subnetworks, or communities from each round, were recursively

subjected to an additional rounds of community detection until finer-scale population structure could no longer be stably detected (Figure 6.1).
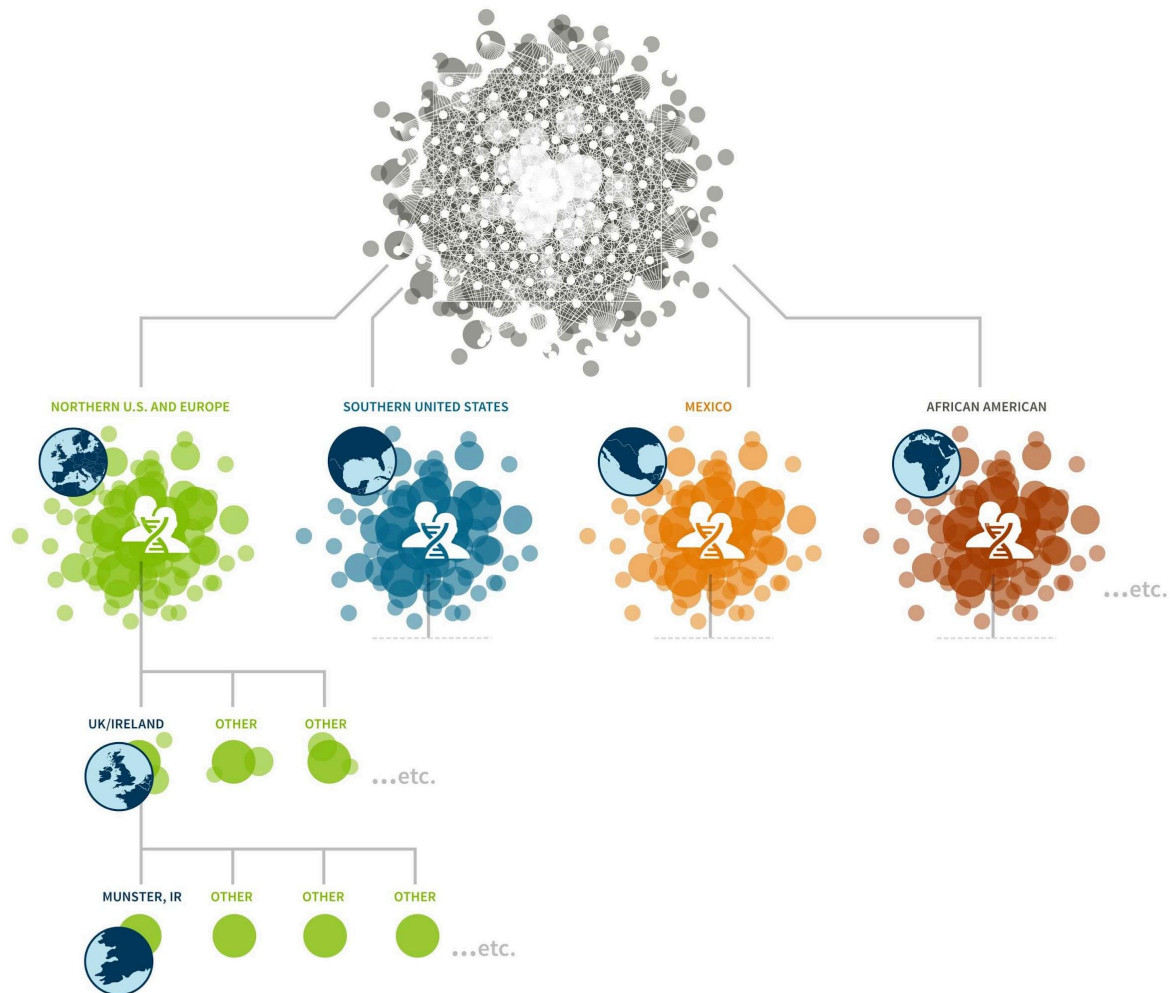

Figure 6.1: Recursive application of the community detection algorithm on subnetworks reveals finer-scale population structure within the larger IBD network.

For example, the first round of community detection could discover a large community of millions of people with ancestry in the Northern US and/or Europe (Figure 6.1). Performing community detection solely on this subnetwork reveals numerous smaller communities that correspond to smaller population groups with more specific histories, when the annotating data are considered. In this case, we are able to break down a larger UK/Ireland community into several smaller Ireland communities representing Munster, Leinster, and eastern Conaught, among others (Figure 6.2).
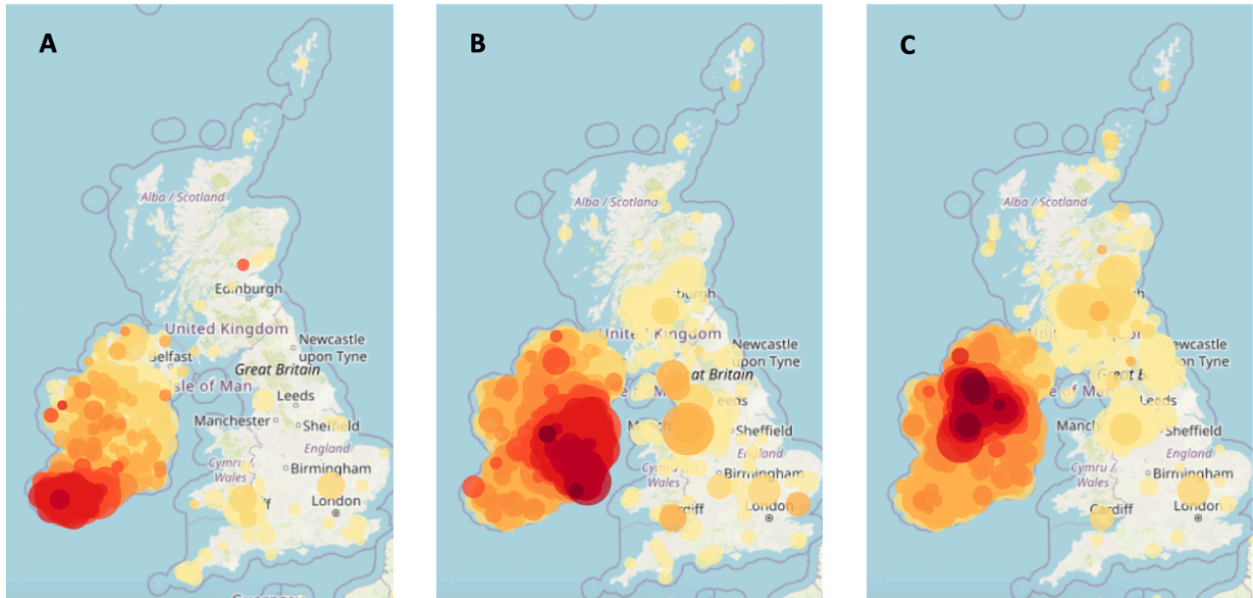
Figure 6.2: Ireland subnetworks discovered through recursive community detection. This figure depicts enriched ancestor birth locations for three journeys originating in Ireland: Munster (A), Leinster (B), and eastern Conaught (C). These and other journeys were identified through community detection on a larger Ireland journey. Significantly enriched birth locations are in red, and less enriched locations are in yellow. Data are projected onto maps from OpenStreetMap (openstreetmap.org/copyright).

The population structure visualized in Figure 6.2, discovered with the same algorithm as before, represents a finer population structure than the one we discover from the entire IBD network. We can continue to run community detection on this smaller set of individuals. As before, we find a number of subnetworks, each corresponding to even finer-scale population structure.

After multiple rounds of recursive community detection, we discover increasingly fine-scale population structure corresponding to local population histories. At the end of multiple rounds of community detection on the *Munster Irish* journey from Figure 6.2A, we discovered over 60 journeys corresponding to several overlapping regions (Figure 6.3).

Figure 6.3: Additional journeys detected in Munster, Ireland. By running community detection on the ancestral journey corresponding to Munster, Ireland, we discover over 60 journeys, illustrated by the white polygons. Data are projected onto maps from OpenStreetMap (openstreetmap.org/copyright)

Considering our DNA database is composed of over 25 million individuals, running community detection on the complete database at a single time is computationally infeasible. Therefore, we have adopted a targeted approach where we focus on starting with DNA samples from individuals with likely ties back to certain global regions. Under this approach, if our goal is to develop additional discoveries in Ireland, we may begin our community detection on the IBD network of individuals already assigned to existing Irish journeys, or filter individuals based on

their ancestral region results. This improves the computational efficiency and performance of the detection process.

7. Assigning Individuals to Ancestral Journeys

While the results from the recursive application of the community detection algorithm on the IBD network reveal intriguing fine-scale subnetworks, we still require a way to deliver these insights to customers. One possibility is that we could select the single community that each sample is assigned to at the end of the community detection algorithm and deliver that as an ancestral journey assignment.

This approach would have two fundamental limitations. First, any single AncestryDNA sample may have strong connections to multiple journeys. For example, an individual with shared ancestry with both an Irish journey and shared ancestry with an Italian journey may have a strong connection to both journeys, but due to the nature of the detection algorithm we use, the end result would only be one assignment. Second, running community detection daily for a large network with millions of samples and billions of connections is computationally infeasible.

Instead, we have opted to use machine learning algorithms, which overcome both of these limitations. To assign samples to ancestral journeys, we create a **reference panel** of samples for each journey that is discovered. Each reference panel is refined to remove individuals less representative of the journey and to account for close family relationships. For each reference panel (representing one journey) that passes certain quality metrics, we construct a **binary classifier** (Figure 7.1).
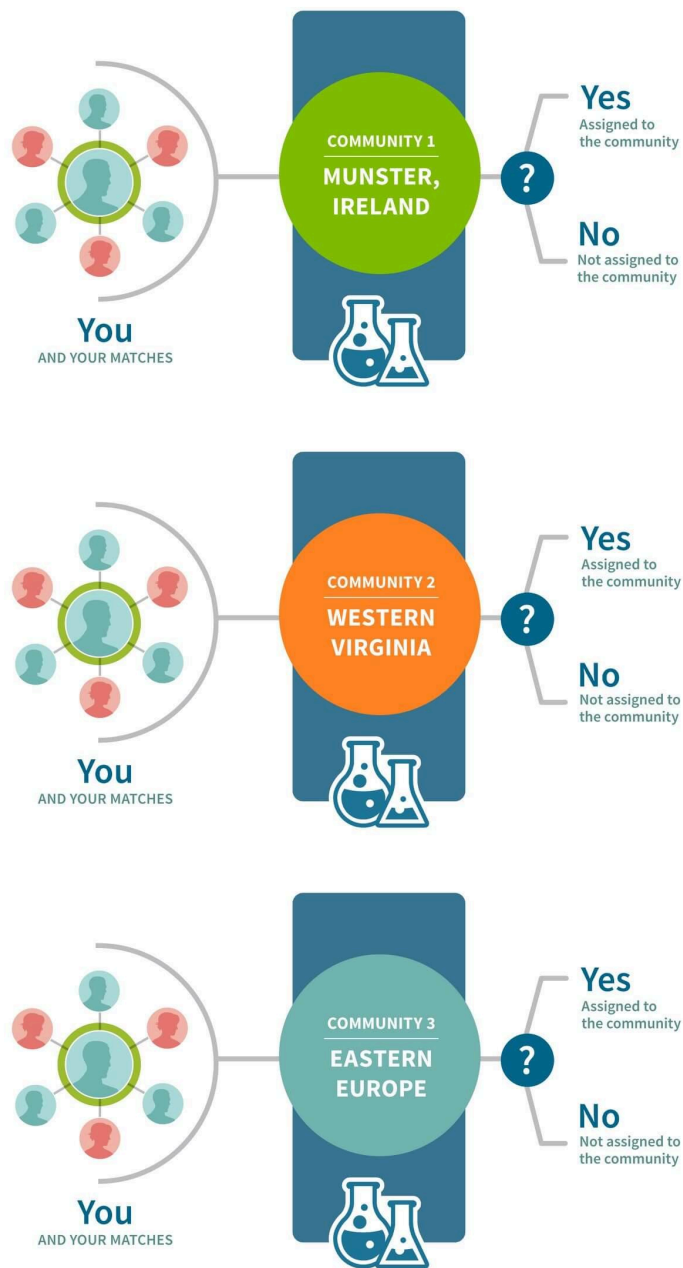
Figure 7.1: Binary classifiers are used to assign AncestryDNA customers to multiple ancestral journeys. For each of the ancestral journeys, we build a binary classifier that will decide if a customer should be assigned to that journey or not.

Binary classification is a machine learning approach that assigns a sample to one of two results given a set of features. For example, given features describing a sample's IBD connection to the network, a classifier will decide "yes, assigned", or "no, not assigned." Since a separate binary

classifier is built for each journey, an individual has the potential to be classified "yes, assigned" for multiple journeys. For example, an individual with shared ancestry from two journeys may be assigned to both.

This approach to journey assignment can be described as a multi-way classification problem, in which each sample may be classified into zero, one, or more journeys (Figure 7.2). Using this multi-way classification scheme, we are able to assign individuals to many journey and side-step the computationally infeasible task of running community detection on the full AncestryDNA database with each new sample.
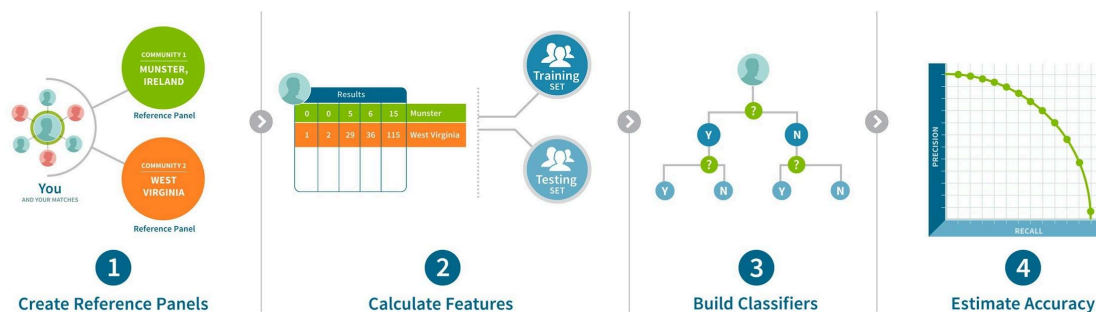
Figure 7.2: An overview of the multi-way classification scheme. (1) For each of the discovered journey we create a reference panel from the highly interconnected DNA network of individuals. (2) A feature vector is created representing the number of close and distant IBD connections with other individuals in the large IBD network of all AncestryDNA customers. (3) Based on those features, the classifiers make a yes/no decision to assign the customer to any of the journey. (4) A validation set is used to estimate the accuracy of each classifier.

The features that are used in these classifiers are found by summarizing each sample's IBD connections in the IBD network and the discovered journey. Because not every generated feature is useful for each classifier, we use standard feature selection techniques to select only the most informative features for each model. The number of selected features varies for each classification model.

For each journey, we use the selected features to train a binary classifier that can be saved and used to assign any AncestryDNA sample to zero, one, or more relevant journey. We identify low, medium and high confidence thresholds for each classifier by maximizing the f-score on the training set.

We use a validation set (a set of samples that were clustered into a community, but held out from the set of samples that were used for training) to estimate the accuracy of each classifier.

8. Determining Parent-of-Origin for Ancestral Journey Assignments

In the preceding sections, we show that we can detect networks of IBD among a large number of genotyped individuals who are part of our database. By leveraging the Louvain community detection algorithm, we identify stable genetic community structure that accurately reflects populations histories. Assigning customers to these journeys provides them a richer family history experience.

However, it remains a challenge to identify through which parent an individual is linked to their ancestral journey. For example, whether an individual has inherited their *Munster Irish* through their mother or through their father. This level of detail is especially useful in genealogical research as individuals attempt to trace their origins back through multiple generations.

Therefore, we designed a novel machine learning approach to infer the inheritance pattern of an individual's journeys. In the example illustrated in Figure 8.1, a customer can see that they inherited their *Southern Utah & Arizona Settlers* from their father specifically, while their other Utah Settler journeys were inherited from both parents. In the following sections we describe our approach for determining the parent-of-origin for a customer's journeys.

Figure 8.1: Example customer results where journeys are separated by parent-of-origin. In their results, customers* can see their journey assignments, which journeys were inherited from each or both parents, and genetic relatives ("Matches") that also share that journey. *This DNA feature requires an Ancestry membership.

8.1. Phasing Customer DNA Data & Separating Matches

A crucial step in determining the parent-of-origin for journey assignments is to separate an individual's DNA into the haves inherited from each parent—a process called **phasing**. Specifically, at every site in a person's DNA, they inherit two versions of the DNA marker, one from each parent.

In order to separate the DNA inherited from each parent for all of an individual's DNA, we utilize our proprietary technology called SideView™. In brief, SideView™ works by comparing the DNA a person shares with their matches and the DNA those matches share with each other. This is based on the premise that all of a person's matches share one or more segments of identical DNA with them and at least one of their parents. As well, a match is usually related to an individual through only one parent.

When IBD segments from different matches overlap but don't match each other, they indicate there is only one way to reasonably phase the person's DNA in that region. By leveraging our DNA database, it is possible to find enough of these overlapping IBD segments across a person's DNA to phase all of their DNA. As a person's DNA is phased, we keep track of which matches group together and share DNA with each other. At the end, we have phased their DNA and organized their matches into three groups: those connected to the paternal lineage, those connected to the maternal lineage, and those that are connected to both parental lineages.

SideView™ uses DNA shared with distant relatives to power the phasing. The correctness of the DNA phasing for a person therefore relies, in part, on that person sharing enough DNA with other people in our database. For certain groups of people, there are fewer matches in our database and so there is a lower level of completeness for their phasing. As well, certain populations have historically high rates of endogamy, which violates our method's assumption that a match is related to a person through only one parent. In this case, any match may be related to both parental lineages, limiting our ability to accurately phase the DNA and organize the matches.

8.2. Building Classifiers Based on Phased Data

Each ancestral journey is characterized by a set of reference panel individuals. We create phased and unphased feature vectors for reference panels by quantifying phased and unphased DNA matches among reference panel individuals, for each genetic journey. Specifically, for each reference panel individual, we determine which half of their DNA–paternal or maternal–has higher matching to the remainder of the reference panel, and keep that half of their DNA in the reference panel. We then perform feature selection and train 2 predictive models (1 for unphased data and 1 for phased data) that can assign a person to a journey.

8.3. Determining Parent-of-Origin for Assigned Journeys

When journey assignment models are developed for phased data, we are able to accurately

determine the parent-of-origin for an inherited journey a large percentage of the time. In a test sample of ~2 million people assigned to French-American ancestral journeys, we identified the parent-of-origin for the journey assignment in 95% of the cases (Figure 8.2).
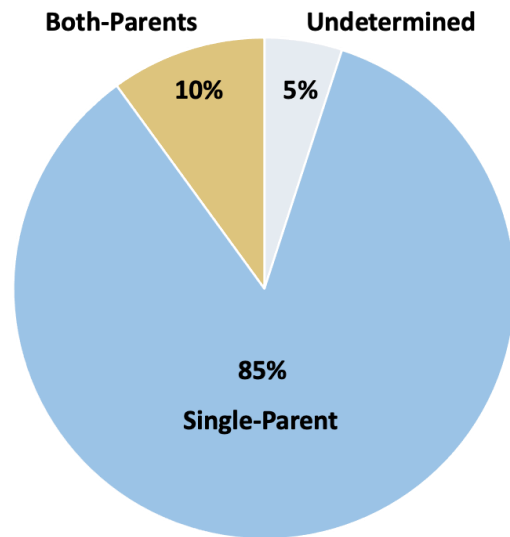

Figure 8.2: Results from identifying the parent-of-origin for the assignment of ~2 million people to French-American ancestral journeys. Overall, we were able to determine the parent-of-origin for 95% of the assignments of several French-American journeys. Approximately 85% of individuals were connected to the journey through DNA they inherited from one parent, while 10% were connected through DNA they inherited from both parents.

In a separate analyses of a subset of our journeys, we examined results for trios of individuals where we had journey assignments for a child and both of their parents. In these cases, we could assess how well the child's phased DNA journey results matched the results for their parents. Specifically, we looked at the proportion of the journeys in the child's phased results (where we infer which journeys are inherited from each parent) that are shared with each of the parents'.

When comparing the child's inferred parental journey assignment with their parent's own assignments, the journey assignment matched more than 95% of the time (Table 8.1). Our ability to make a parent-of-origin assignment was correlated with our ability to accurately and completely phase a person's DNA.

	Parental Designation Concordance Rate					
Journey Group	**0.00 - 0.25**	**0.25 - 0.50**	**0.50 - 0.75**	**0.75 - 0.99**	**0.99-1.00**	**Overall Average**
Southern European	617	854	983	102	115766 (**97.8%**)	0.988
Middle East and North African	88	106	75	0	14816 (**98.2%**)	0.988
French-American	703	884	1524	266	74283 (**95.6%**)	0.978

Tabel 8.1: Concordance rate (binned) between a child's phased journey assignment and their parent's unphased journey assignments.

9. Conclusion

In this white paper, we describe our Genetic Communities™ technology, which identifies and assigns individuals to communities of other customers with whom they share more DNA matches than with the rest of the database. These communities correspond to ancestral journeys defined by fine-scale population structure due to very recent, and often documented, historical patterns.

First, we identify identical-by-descent (IBD) genetic connections among millions of AncestryDNA samples. When these connections are aggregated into a network, our computational methods reveal densely-connected clusters, which we refer to as communities. Members of each community are more related to each other than to members of other communities. Next, using user-generated pedigrees and inferred ancestral regions, we infer the ancestral journeys of these communities, identifying the likely historical origins of the population substructure and patterns of population movement and settlement. Finally, by applying machine learning techniques, we infer membership of AncestryDNA customers to these ancestral journeys, thereby providing a detailed survey of their contemporary family history around the globe.

As the AncestryDNA database continues to grow, we expect that our ability to discover additional structure in the IBD network will be enhanced. This will likely lead to discoveries of ancestral journeys in new areas of the world and with enhanced granularity, leading to a richer family history experience for all AncestryDNA customers.

10. References

Blondel, V. D., Guillaume, J., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory E*, P10008 (2008).

Csardi, G., Nepusz, T. The igraph software package for complex network research. *Complex Systems*, 1695 (2006).

Fitzgerald, Patrick and Brian Lambkin. *Migration in Irish History 1607-2007*. Palgrave McMillan, 2008.

Laidley, W. S. *History of Charleston and Kanawha County, WV, and Representative Citizens*. Chicago: Richmond-Arnold Publishing Co., 1911.

Noto, K., Ruiz, L. Accurate genome-wide phasing from IBD data. *BMC Bioinform.* 52 (2022).

Rice, Otis. *West Virginia: A History.* Lexington: The University Press of Kentucky, 1993. https://muse.jhu.edu/ (accessed May 13, 2105).