Estimation of Recent Ancestral Origins of Individuals on a Large Scale

Ross E Curtis
AncestryDNA
1300 W Traverse Parkway
Lehi, UT 84043
rcurtis@ancestry.com

Ahna R Girshick
AncestryDNA
153 Townsend St, Ste. 800
San Francisco, CA 94107
agirshick@ancestry.com

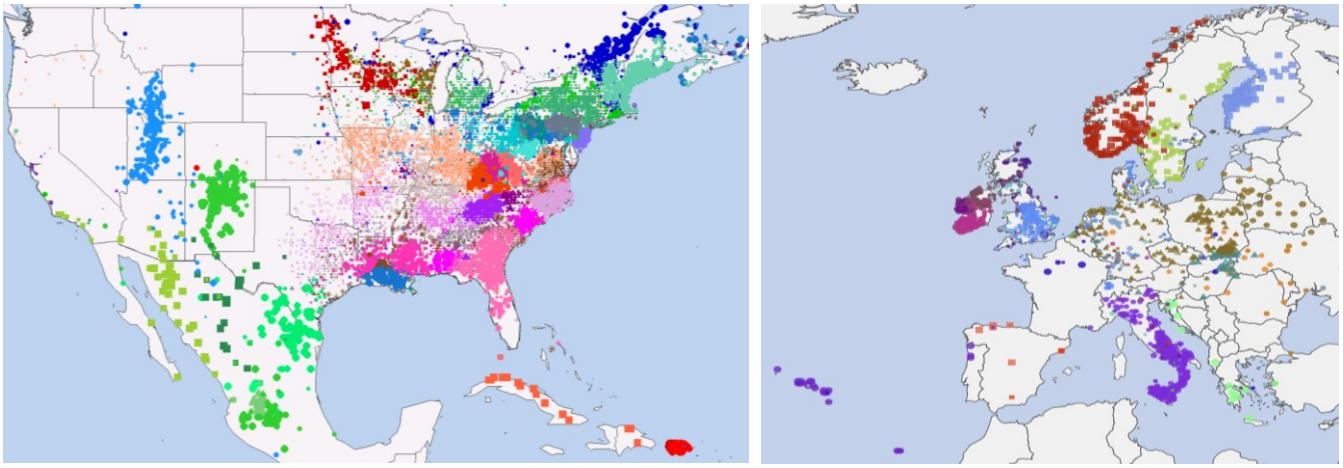

Figure 1. Ancestral birth locations, or *community enrichment regions*, of 60 discovered genetic communities. The points on this map represent enriched ancestral birth locations of sub-networks of individuals who were discovered in a large genetic network. Using machine learning and genetics, we demonstrate the ability to accurately map new individuals to these subnetworks, thus discovering their recent ancestral origins. Three regions located in South America and the Pacific are not shown. The shape and color of the ancestral birth locations were selected for visual aesthetics.

ABSTRACT

The last ten years have seen an exponential growth of direct-to-consumer genomics. One popular feature of these tests is the report of a distant ancestral inference profile—a breakdown of the regions of the world where the test-taker's ancestors may have lived. While current methods and products generally focus on the more distant past (e.g., thousands of years ago), we have recently demonstrated that by leveraging network analysis tools such as community detection, more recent ancestry can be identified. However, using a network analysis tool like community detection on a large network with potentially millions of nodes is not feasible in a live production environment where hundreds or thousands of new genotypes are processed every day. In this study, we describe a classification method that leverages network features to assign individuals to communities in a large network corresponding to recent ancestry. We recently launched a beta version of this research as a new product feature at AncestryDNA.

1 INTRODUCTION

Direct-to-consumer (DTC) genomics is transforming the way individuals see themselves, do genealogical research and think about health and wellness [23]. The DTC genomics industry started accelerating in 2007 and has witnessed exponential growth. At the end of 2016, more than 4.5 million combined genotypes were reported for the top three largest DTC genomics companies, with over three million genotypes at AncestryDNA alone [2]. At AncestryDNA, with a sample of saliva, individuals receive an estimate of the distant ancestral origins of their DNA, connect with close and distant family members, and discover new ancestors.

DTC genomics testing can provide a powerful experience for individuals wanting to learn more about their past. For example, a DNA test can reveal to an individual, with generally high confidence, that his or her distant ancestry is approximately 40% Eastern European and 25% Scandinavian (Figure 2), thus revealing intriguing details about one's identity and family history. This type of profile generally uses selected ancestry-informative *single-nucleotide polymorphisms* (genetic markers with variation at a single locus, also referred to as SNPs) distributed across the genome to estimate the proportion of an individual's genome that has originated from different populations around the world.

KDD '17, August 13-17, 2017, Halifax, NS, Canada

ACM ISBN 978-1-4503-4887-4/17/08.
http://dx.doi.org/10.1145/3097983.3098042

Methods to do this vary, but include statistical modeling approaches [1] and approaches that leverage hidden Markov models to assign local ancestry to tracts of DNA sequence, or *haplotypes*, across the genome [17, 15].

Distant ancestral inferences are based on the identification of *population structure*, groups of individuals that can be separated, broadly or even subtly, by genetics due to historical mating barriers between groups. One limitation of deep ancestral inference approaches is that the estimate of an individual's ancestry often reflects the distant past, perhaps even thousands of years ago. Yet for many individuals, their more recent ancestral history may be more relevant or interesting. However, due to the reliance on common SNPs, which are generally old, it is difficult for these types of approaches to leverage population structure that may be more recent, say in the post-Columbian time frame of European Immigration to the Americas [11, 19].

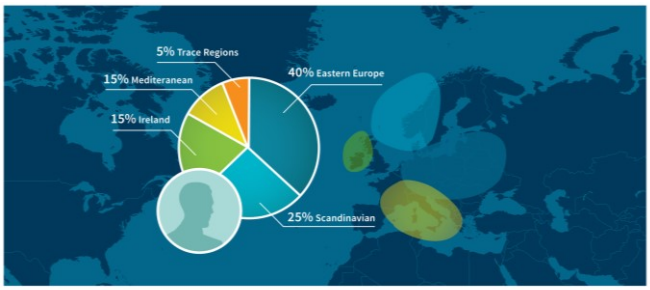

Figure 2. An example distant ancestral inference profile.

Recent work in the population genetics community has identified more recent population structure by analyzing data from individuals with a manually curated history in a single region (e.g., manual research to establish generational stability in a specific geographic region) [16, 18]. These studies have been successful in identifying fine-scale population structure associated with the more recent past, but have not been successful in building out a generalizable method to find structure without manually curating individuals based on geography and pedigrees. Recently, we developed a method capable of identifying much more recent fine-scale population structure than was previously possible using general methods [13]. We have also shown that this fine-scale structure can be linked to a modern understanding of geography, culture, and historical migrations with surprising clarity, such as the descendants of the early settlers in western North Carolina. The structure we discovered is considerably more recent in time than that of most population genetics research, generally within the last few hundred years.

Our technique uses community detection methods and is described more fully in Section 3. By demonstrating that fine-scale structure can be discovered in a network without manually curating individuals *a priori*, we have shown that it is possible to associate an individual with recent populations from which he or she may have ancestry, possibly bringing to light more details about their recent past. For example, if we consider our example from Figure 2, in addition to discovering the broad distant-ancestry regions in Europe, an individual may find out that they belong to a small community of people who have recent shared ancestry in Southern Poland and another small community with ancestry from Western Norway. In our research, we found communities world-wide including in Europe, the Americas, and Asia. Some examples include a community of individuals who share European ancestry in the Appalachian Mountains of Western Virginia or a community who share ancestry from African Americans in 19th century South Carolina. We can refer to these discovered subnetworks as *Genetic Communities*™.

1.1 Challenges of Community Detection in a Commercial Product

While we have previously demonstrated the ability to discover these rich, diverse, and historically relevant community structures [13], it is not intuitively obvious how to deliver these results to individuals as part in a rapidly growing database. There are four main limitations of the approach we have taken to discover communities that are not desirable for the efficient assignment of new individuals to one or more genetic communities.

1.1.1 Scalability

The potential to discover new genetic communities increases as Ancestry DNA's user base grows and becomes more diverse, making the product more attractive. However, community detection is computationally challenging, and the computational challenges increase as the network grows. Although our graph is sparse, it has billions of edges, the number of which is growing quadratically. Running community detection on a daily basis would require a significant dedication of computational resources. Individuals generally expect results to be available as soon as possible after they send in their DNA sample, which may be difficult if required to perform community detection in a large, growing network.

1.1.2 Multiplicity of Assignments

Many community detection methods that scale up to millions of nodes and billions of edges assign each node to the "best" cluster, thus missing opportunities to assign an individual to multiple secondary populations if they are all relevant (e.g., the individual from Figure 2 may have parents from different populations and thus might expect to be assigned to a more specific Scandinavian as well as an Eastern European community, instead of just one). For assignments to be relevant across multiple ancestries for an individual, we want the flexibility to assign an individual to as many relevant communities as is possible while maintaining high precision.

1.1.3 Consistency of Results

Many individuals expect genetic results to be deterministic, and they want to interact with other members who are assigned to the same genetic community. Exploring a genetic community gives insights to an individual about their recent ancestry from the shared geographic or cultural region. However, community detection methods are stochastic in nature, and are highly dependent on the overall network structure which is changing as new individuals are added. Even if the same structures are observed consistently, individual assignments for a particular individual may not always be consistent across repetitions.

1.1.4 Interpretation of Communities

The interpretation of communities using a non-manually curated set of individuals presents a significant challenge in understanding the geography and historical forces that contributed to the formation of the community. Given a stable set of pedigrees and individuals, historians can augment the user experience by researching the historical context of the mating

barriers and events that may have led to the formation of the community in the network. For example, a community in Western Virginia can be enhanced with stories about the original settlers of Western Virginia, how they migrated from Europe to the United States, and what forces kept them together over the course of many generations. To appropriately research and understand these histories, the structure of the discovered communities needs to remain stable while deployed in the product.

1.2 Classification Method Addressing Challenges

To overcome these significant challenges, we turn to machine learning classification to assign individuals to predefined communities discovered through community detection. Our solution uses community detection results as training data that is used to build classifiers based on connections in a large network of genotypes. We build a machine learning classifier for each community. This enables an individual to be considered for hundreds or even thousands of communities with very few computational resources, enabling assignments to relevant communities in the product experience. A version of this system is deployed as part of the AncestryDNA product.

This paper proceeds as follows. In Section 2, we introduce a basic description of selected genetics concepts that we leverage to construct a large network of individuals' genotypes, and general background on community detection methods. In Section 3, we describe our approach to use community detection to find structure in a genetic network. Section 4 details how we create features, prepare training data, train, test and validate the classifiers. We close with a discussion in Section 5 and conclude in Section 6.

2 BACKGROUND

In this section, we first describe some basic genetics principles that we apply to construct a genetic network for discovering fine-scale population structure. Then, we detail how that network is created. We also discuss prior work in community detection and some of its applications.

2.1 Distant Ancestral Inference

All humans have DNA, or a *genome*, which is passed down from parent to child over each generation. Thus, our genome is a mosaic of fragments of the genomes of our ancestors and can tell us about where and who we came from.

One can think of the human genome as a string of about three billion characters (nucleotides made up of A, T, G, or C). When a genome has one character at a particular position and another genome has a different character at the same position, we call this a single-nucleotide polymorphism, or a SNP. A *haplotype* is a unique sequence in a genome which is inherited as a unit from a specific ancestor.

Variation represents changes in our genome that happened at random times and places and then moved with human populations as we populated the globe [20]. As humans moved from Africa to Europe, Asia, and the Americas, groups split apart, taking with them their genetic variation. By chance, the genetic variation of groups settling one area could be different than those that settled in another. Over time, due to the stochastic nature of genetic inheritance through the generations, some versions of a SNP might become much more common in one population than in another. Differences in SNP frequencies between populations can increase over time, especially if the populations have been limited to small geographic areas and individuals from them have generally mated exclusively within each population.

At AncestryDNA, we use high-throughput SNP-chip technology that allows us to assay variation at hundreds of thousands of SNPs for relatively low cost [14]. Any particular SNP is not usually informative on its own, but jointly looking at hundreds of thousands of SNPs can provide an accurate distant ancestral inference (reflecting global population substructure). Using the basic assumption that SNP frequencies differ between populations, we use a statistical model called ADMIXTURE [1] to determine an ancestral inference estimate for each individual. ADMIXTURE is a fast optimization of a model first developed in STRUCTURE [24] to find a distant ancestry estimate for each individual in consideration. The model assumes that there are a specified number of unique populations (or "distant-ancestry regions"), each with a different frequency at each SNP considered.

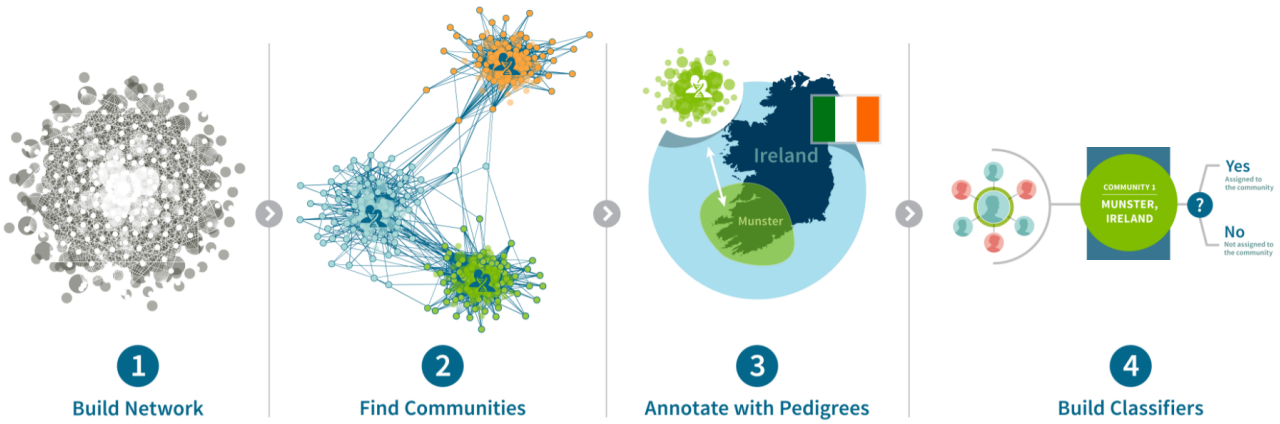

Figure 3. Flow of creating genetic communities and community assignment classifiers. 1) We build the network based on the IBD matches between individuals (Section 2.2). 2) We use the Louvain method for community detection to discover community structure in the large IBD network (Section 2.3). 3) We use public pedigree information aggregated by community to understand the common history of the community members (Section 3.2). 4) We build classifiers to assign users to communities based on selected features (Section 4).

Each individual is assigned a weight vector with an entry for each population. The vector can be thought of as a probability vector where each entry in the vector describes the proportion of the individual's genome that comes from that population. A matrix of these distant ancestry estimates over all individuals is estimated using a block-relaxation approach and maximum likelihood optimization. The result is a vector for each individual that represents their distant ancestral estimate—the amount of that individual's genetic variation that originates from each of the selected global regions.

2.2 Building a Genetic Network

While variation across multiple positions in an individual's genome provides a portrait of the more distant past (e.g., hundreds to thousands of years ago), haplotype variation provides information about the people an individual connects to more recently (e.g., in the past hundreds of years). Shared haplotypes represent a relationship between two individuals through a recent shared ancestor, and the amount and length of shared haplotypes is proportional to their distance in relationship. Generally, the longer the haplotype, the more recent the shared ancestor. For example, two individuals who share half of their genome are a parent and a child while two individuals who share a short segment of length 18 centimorgans (a unit of measurement for genetic length, abbreviated as cM) are generally estimated to be 3rd to 6th cousins.

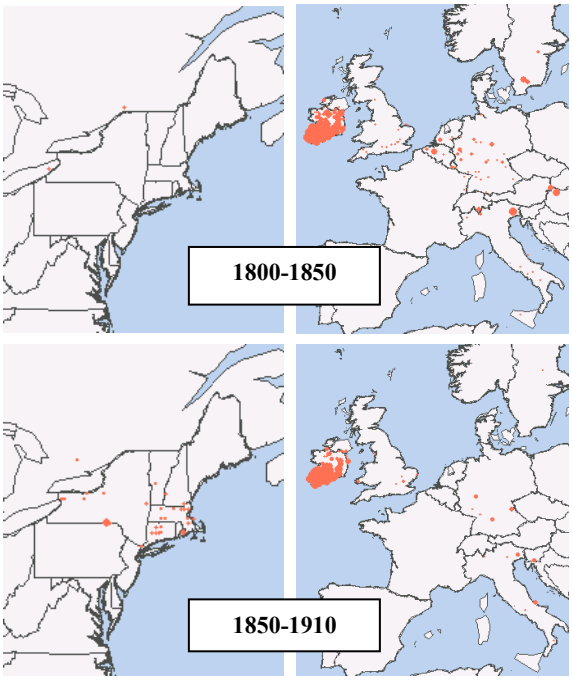

Figure 4. An example of enriched birth locations for a community. In this case, two time frames are considered for a community. We observe ancestors in Munster, Ireland in both time frames, suggesting that the community members have common ancestry in Munster. In the second time frame, we observe enriched birth locations in the New York and Boston areas, consistent with the historical migration of the Irish to the US in the 19th century. In this figure, birth locations with an odds ratio greater than 5 are shown. Size is determined by the frequency of the birth location in the aggregated data. Figure 1 was similarly created using only the enriched locations for 60 communities in the 1850-1910 timeframe.

Haplotype estimation, or phasing, is the process of statistically estimating the haplotypes from the SNPs of an individual and a population [5, 22]. Given phased haplotypes, an *identity by descent* (IBD) analysis is a technique for finding shared haplotypes [12]. Given the discovered haplotypes, other methods are used to estimate the relationship between two individuals.

To create a genetic network based on IBD, we construct a weighted network of relationships between individuals where the nodes represent individuals and the edges represent identified IBD relationships between them. These edges can be weighted by the length of the IBD, and are filtered to only include segments longer than 12cM. By excluding short segments, the relationships between individuals in the network generally represent more recent shared ancestry between individuals.

2.3 Community Detection Methods

Community detection algorithms are network clustering algorithms that identify strongly connected subsets of a network. They have been used in many applications including those in sociology, physics, biology, and computer science [9]. There are many available methods for inferring communities from network data [20]. As we discuss in Section 3, in [13], we employed the widely-used Louvain method [4]. It is known for its simplicity and ability to perform quickly on large networks. Like other community detection methods, it attempts to optimize the "modularity" of a partition of the network. The modularity measures the density of links inside communities compared to links between communities on a scale of -1 to 1 [10]. The optimization first looks for small communities by optimizing modularity locally. Then it aggregates the nodes belonging to the same community and builds a new network whose nodes are the communities. These two steps are repeated iteratively until a maximum of modularity is attained, thus discovering a community structure.

3 RECENT POPULATION STRUCTURE

In this section, we describe our recent work in discovering structure within a genetic network using community detection [13]. In both that research and this paper, we use the same dataset of genotypes from 742,394 individuals who have consented to participate in scientific research. In this section we describe how we discovered significant structures within this network using community detection.

3.1 Community Detection in IBD Networks

We first created a large genetic network with the 742,394 individuals as nodes as described in Section 2.2. Next, we used community detection methods to discover structure in this network. Intuitively, because estimated IBD connections between individuals are likely due to recent shared ancestry (within the past 10 generations), clustering patterns in this large network likely represent recent shared history. The result is that we can identify communities of living individuals that share large amounts of DNA due to specific, recent shared history.

We used the Louvain Method as implemented in the *iGraph* package [7] to discover these communities. In our case of an IBD network, communities represent groups of individuals that are

more related to one another than they are to others in the large IBD network.

3.2 Interpreting Communities Using Metadata

The genetic network and community detection steps are based solely on genetic data, which is not easily interpretable. To interpret the communities that we discover, we use data annotation, which has the power to supplement an individual's genetic data with metadata. The two pieces of annotating metadata we use are 1) individual pedigrees and 2) distant ancestral inference. We will use the example community of Munster, Ireland to illustrate how these data are used.

3.2.1 Aggregated Pedigrees

In addition to offering a DNA test, Ancestry also offers a genealogy product in which individuals can create a digital family tree, or *pedigree*, with an optional link to their DNA test. A family tree is composed of the parent/child relationships, can extend multiple generations, and can include information such as birth date, birth place, or surname for any number of ancestors for an individual. We use the pedigree data, in aggregate, to interpret the genetic communities identified from community detection.

We aggregate the pedigrees of the individuals who belong to each identified community, and filter based on temporal frames. The next step is to calculate the odds ratios for locations and surnames to find those that are enriched to a particular community in a time period of interest.

By stratifying enriched birth locations by time frame, we can observe the migration of a community over time and identify significant regions where many of the community's ancestors lived (Figure 4). In this example, the ancestors of the community are enriched for birth locations in Munster, Ireland. After 1850, we start to see effects of the Great Famine and the subsequent migration to the United States in Boston and New York City. The top enriched surnames that we observe for ancestors in this group

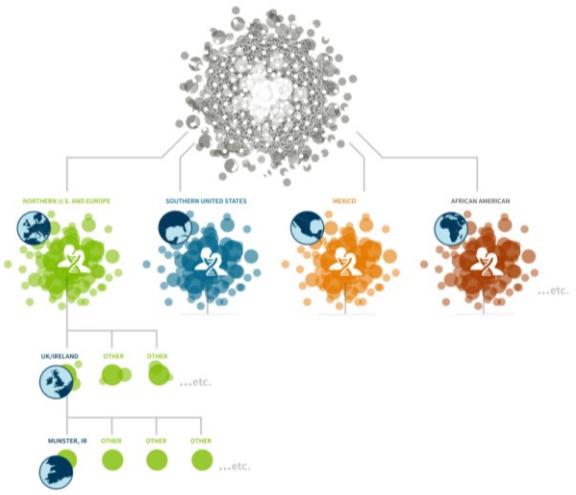

Figure 5. An example of how high-resolution communities are discovered with iterative community detection. We start with the entire network, which includes 742,394 individuals. After the first round of community detection, several sub-communities are discovered. We run community detection iteratively on these sub-communities to find higher-resolution communities, such as the community in Munster, Ireland.

are Sullivan, O'Brien, McCarthy, and O'Connor, which are historically Munster, Irish surnames.

3.2.2 Distant Ancestral Inference Information

Another piece of annotating data that can be considered for interpreting each discovered community is the distant ancestral inference data that we generate from each genotype as explained in Section 2.1. For example, for the community that we have been considering with birth locations in Munster, Ireland, we found that their ancestry is primarily from Ireland (more than 95% have at least some estimated ancestry from Ireland).

3.3 Iterating to Find Fine-Scale Structure

When we run community detection on this large network of 742,394 individuals, we find six top-level community structures. These six communities are: African and European Americans in the US South, Europeans in Europe and the US North, Mexicans and South Americans, Caribbeans, European Jewish, and Asians [13].

While these groups are interesting, they are coarse in scale and could have been discovered by an ancestral inference algorithm itself such as ADMIXTURE, mentioned above. However, it is possible to find higher-resolution communities through the iterative application of the community detection algorithm. Since each observed community is itself an IBD network on which we can apply the same community detection algorithm to discover sub-communities, we performed community detection recursively.

As an example, consider the Munster Irish discussed in the previous section (Figure 5). The first round of community detection discovers a large genetic community comprised of hundreds of thousands of people with ancestry in the Northern US and/or Europe. Performing community detection solely on this subnetwork reveals several smaller genetic communities that correspond to smaller population groups with more specific histories. In this case, we find genetic communities representing individuals with ancestry in Italy, Pennsylvania, New York, and the UK and Ireland. By iteratively applying this technique on the UK and Ireland sub-community, we find higher-resolution population structure. We find three communities that have ancestors from Ireland, one of which is from Munster, Ireland.

In any network, it may be possible to find a set of clusters with the Louvain method for community detection, whether the structure is meaningful or not. Our goal is to identify reproducible subnetworks, or genetic communities. In other words, we seek to identify communities that appear regardless of the specific individuals used to construct the initial network, as these communities are more likely to correspond to real genetic structure. Therefore, we perform bootstrap experiments to measure the extent to which each internal cluster in the hierarchy produces the same subclusters in community detection. To do so, we replace each internal network with a random subset of 80% of its nodes and rerun the Louvain method. Then we investigate whether we observe the same community structure as we do in the full network. If the structure is consistent and modularity improves in subsequent iterations, we continue to iterate. For this study, we go through three rounds of community detection. This process results in 63 genetic communities. For more details about this method and accompanying results, please see our previously published work [13]. In the following sections, we build on this methodology to assign individuals to discovered communities.

4 COMMUNITY ASSIGNMENT CLASSIFICATION

The community detection algorithm requires a few days to run and is not efficient to completely rerun for each new individual. Furthermore, the Louvain method only assigns one community per individual. We create community assignment classifiers for rapid assignment of each new individual to their appropriate genetic communities as they come in, one at a time. We seek to allow individuals to be assigned to multiple communities, when appropriate. By building multiple independent classifiers, we are able to assign each individual to zero or more relevant communities continuously, without affecting the assignments of other individuals. We explain our technique to determine community assignments below.

4.1 Stability Experiments

An outcome of the community detection refinement process described in Section 3.3 is that we can observe the frequency that individuals are assigned to different communities. Thus this analysis is a way for us to assess not only the stability of the discovered clusters, but also the stability of each individual in a particular cluster. Because we desire the training data to contain high-quality, unambiguous labels, we use the stability of an individual's presence in a community as an indicator of label quality.

For the stability experiments, we suppose that we have N individuals in a given cluster in the community structure. We build an IBD network composed of those N individuals. Suppose that we detect C communities A_1, A_2, ..., A_C by the Louvain method. Next, we create a new bootstrap set by randomly choosing 0.8×N individuals. We use these individuals to build an IBD network and run community detection using the Louvain method. Suppose that we detect D communities from this subnetwork: B_1, B_2, ..., B_D.

Now, we consider each community B_j (j=1,..,D), and find the most consistent community A_i (i=1,...,C). We define the most consistent community A_i with respect to community B_j as a community with the largest overlap between communities A_i and B_j, given by the *Jaccard Index*, that is, the ratio between the sizes of the intersection and the union of two communities ($|A_i \cap B_j| / |A_i \cup B_j|$). In this way, we associate each newly discovered community B_j with an original community A_i. If multiple clusters in A have the same or a similar (within a predefined threshold) Jaccard Index, B_j will be assigned to each of them. That is, we find a one-to-one or one-to-many mapping from B_j to A.

We do the above for each of 20 bootstrapping runs (increasing the number beyond 20 runs did not significantly change the results). For each individual for each run in which they are included, this results in an assignment to zero or more of the original communities.

We use these results to determine individual stability to a community. For each individual assignment to a community, we calculate the stability score S of that individual's assignment to that community by summing the number of times they are assigned to that community overall, divided by the number of times that they are included in the bootstrapping set. As an example, an individual may be originally assigned to community A_x. However, in the four bootstrapping runs that she is included in, she is assigned to a community B_j that maps to A_x in run 1 and run 4 (e.g., B_j and A_x are likely the same community) to A_y in run 3, and to a community that maps to both A_x and A_y in run 2. Thus, her stability score S is 2.5 / 4 = 62.5% for community A_x and 1.5 / 4 = 37.5% for community A_y.

4.2 Creating Labels and a Training Dataset

In contrast to the unsupervised nature of the community detection algorithm, the classifiers are supervised and thus require accurate labels for training. With respect to any particular community, an individual's label may be positive (member of the community) or negative (non-member part of the community). To determine labels, we use the stability experiment results (Section 4.1). Any individual who is assigned to a community with stability score $S \geq 50\%$ is labelled as a "positive" member of that community. Any individual who is never assigned to that community ($S = 0\%$) is labelled as a "negative" member of that community (e.g., a non-member). Using this method, we find an assignment for 666,083 of the original 742,394 individuals. For the individuals whose stability score S is greater than 50%, the lowest observed stability score is 51%, the 1st quartile is 68%, and the median is 100%. By removing ambiguous individuals ($0\% < S < 50\%$), we build a training set that reflects only those individuals who are most likely to be strongly connected to a community and ignore those individuals who may have ancestry from multiple closely related communities.

Another filtering step that we take to avoid bias in our training data is to exclude all relationships between closely related individuals. For example, a father and daughter will share many of the same IBD connections, potentially biasing the classifier to specific close-family relationships. To remove these relationships, we consider all pairwise relationships where the IBD relationships is predicted to be a parent/child or grandparent/grandchild relationship. For each relationship where both individuals have been assigned to the same dataset, we randomly select one individual to exclude from the training dataset. This results in the exclusion of another 61,843 genotypes to yield a set of 604,240 genotypes. After these two filtering steps, only two communities gain members (due to differences in the community detection results across bootstrapping runs). The average reduction per community is 18%, and only three communities lose more than half of their members.

4.3 Generating and Selecting Features

To create features from our dataset, we used the strength and number of connections that an individual has to the different communities in the genetic network which determine how likely the individual is to be connected to any particular community. Complicating the problem, certain communities are related. For example, it is common for us to see people who have ancestry from both Ireland and Italy, perhaps because many Catholic immigrants from these two populations married upon arrival to the United States. A person who has ancestry only from Ireland may have many IBD relationships with people who are assigned to an Italian community because of this trend. However, the relative number of connections and position in the IBD network can elucidate the individual's strongest connection.

We assume no prior knowledge about the relationships between communities and which communities may or may not be relevant to a particular individual. Instead, for each individual, we create all the possible features and then use a feature selection step to select the most informative features for each particular community classifier.

For each individual in the filtered dataset from above, we generate features by counting the number of IBD relationships that an individual has to the discovered members of each of the 63 communities in the dataset. Each feature is defined as a count of the number of connections between a given individual and all individuals in the filtered list of individuals for a particular community. We separate these counts into five buckets based on the degree of relatedness, that is, by the number of cM shared between the genotypes of two individuals (the more cM shared, the closer the relationship between the individuals). Thus, close family relationships between an individual and a particular filtered community are counted as one bucket, and distant cousins are another bucket. The five cutoffs we use are 12cM (very distant cousins), 18cM (5th-8th cousins), 30cM (4th cousins), 60cM (3rd cousins 1x removed) and 90cM (3rd cousins), which correspond to previous experiments [3] studying the amount of DNA shared by people with varying degrees of relatedness. Given five relationship buckets and 63 communities, we generate 63 x 5 = 315 features for each individual. By bucketing the matches in this way, we are able to account for the strength of the relationships (and thus the time to recent ancestors) of the individual to the community.

Given the diversity of genetic communities and related complex structure of the IBD network, most of the features have value zero because the IBD network itself is sparse. We perform a feature selection step for each community to avoid overfitting, as well as to reduce the computational complexity of training each classifier. For each community, we select the 30 features with the top chi-squared scores between the feature and the labels, weeding out features likely to be independent of the community. Each community thus has its specific set of features to use in building out the classifier.

4.4 Classifier Training

Many classification models may be appropriate for this problem. We use random forest classifiers but believe other models would yield similar performance. We train 63 independent classifiers, one for each community (i.e., one vs. rest strategy).

The number of member training examples per community ranges from 625 (0.10%) to 59010 (9.8%), with a median of 7959 (1.3%). Thus, all community training sets have a substantial class imbalance (e.g., generally about 99% non-members and 1% members). We correct for this by adjusting class weights inversely proportional to the class frequency in addition to assessing an F1 score (a combination of precision and recall) instead of accuracy [6]. This correction helped the smaller communities more than the larger communities.

For each classifier, we perform cross validation using randomized shuffle splits: Ten times we train on a random 60% of the data (*training sets*) and make predictions on the remaining held-out 40% of the data (*validation sets*). We then measure precision, recall, and F1 score of all the predictions of all of the validation data [8]. As another precaution against the class imbalance issue, we stratify the sampling to ensure an equal number of member examples in each training set.

We take two precautions to guard against overfitting. First, we used a large number (300) of decision trees in our random forest. In an initial pilot, 300 trees was an optimal balance of speed and performance. Second, we examined the distribution of the validation F1 scores. We performed a two-sample Kolmogorov-Smirnov test, and based on the p-values greater than 0.01, concluded that 99.7% of the time this distribution was indistinguishable from a normal distribution. This normality suggests that most models were not subject to spurious lucky or unlucky performance as a result of the random draw of the training data.

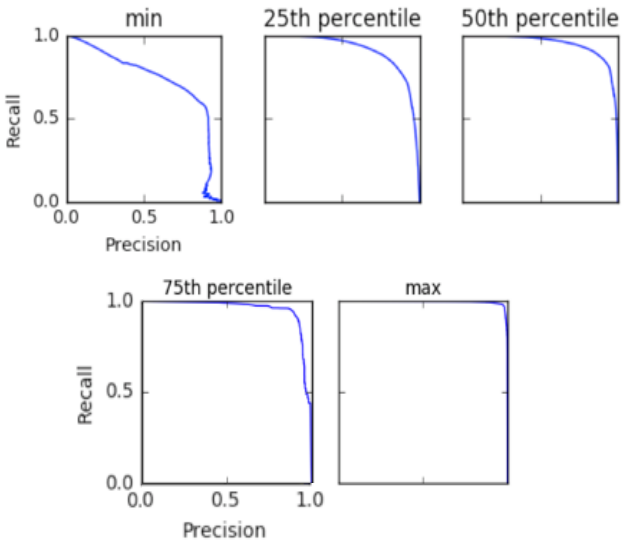

Figure 6. Precision recall curves for the validation data. Five examples are shown representing that with the lowest validation F1 score, the 25th, 50th, and 75th percentiles, and the highest F1 validation score.

This procedure takes about one to two hours to train 63 classifiers in parallel. After training, new individuals can be assigned to communities in real-time (less than a second per individual to create the feature vector from precomputed matches and make 63 assignments).

4.5 Classifier Results and Independent Pedigree Concordance Validation

Using the training dataset developed from the community detection algorithm and stability experiments, we produced 63 machine learning classifiers to be used to assign new individuals to one or more communities. The validation set F1 scores capture the precision-recall performance of each classifier, and are summarized in Figure 6. Representative precision-recall curves representing the minimum, 25th, 50th, and 75th percentiles, and maximum are shown in Figure 6. The median F1 score was 0.88.

To validate the community assignments of these classifiers, we leveraged an independent pedigree dataset. The pedigree dataset contains family trees, dates, and birth locations. We randomly sampled 60% of the individuals in our filtered community detection results dataset (Section 4.2) to create a new training set for this exercise, using stability score S to assign training individuals to communities as before. For each training set, we calculate the odds ratio (OR) of each observed ancestral birth latitude and longitude against the rest of this training set after rounding to the nearest latitude/longitude value. This analysis is the same OR calculation as in Figures 1 and 4, but limited to 60% of the data. We define each *community enrichment region* as the set of all rounded latitude-longitude grid points with an OR of at least five, indicating a geographic area with an enrichment of

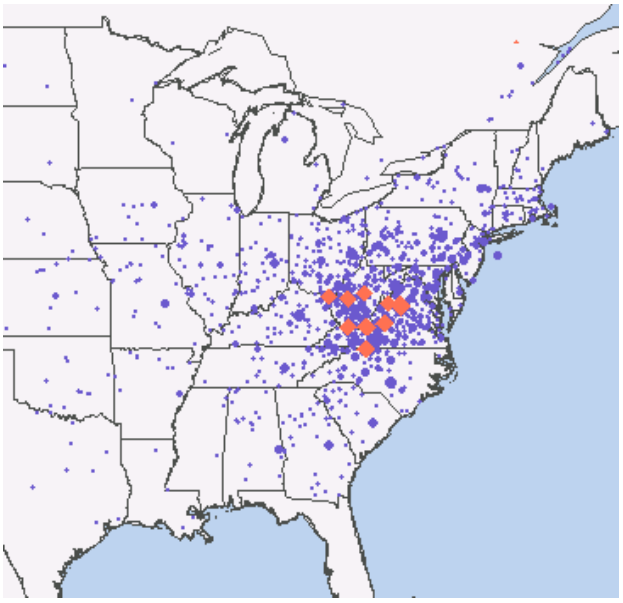

Figure 7. Pedigree Concordance Analysis. We consider pedigree information for each community. This example is around a community made of individuals with ancestry in western Virginia. The orange diamonds on the map indicate the community enrichment regions representing the ancestral birth locations of community members. The purple dots represent all the aggregated ancestral birth locations of test-set individuals assigned by the classifier to the community (not used to create the enrichment regions). We find that 95% of the individuals assigned to this community have at least one annotated ancestor within the orange target region.

birth locations. The orange diamonds in Figure 7 show these grid points for the community enrichment regions for individuals in a community with ancestry from western Virginia. Then for each community, we randomly select 100 community members from the test set (remaining 40% of the data) who have a pedigree with at least eight individuals. We do not consider the shape of the pedigree (e.g., the relationship of the eight ancestors to the community member). We examine the birth locations of all of the ancestors of the 100 individuals (Figure 7, purple dots). We calculate the *pedigree concordance score* as percentage of these community members with at least one of their ancestors born in a location with a latitude and longitude that rounds to one of the target enrichment points (Figure 8). The average concordance score is 91%, and the median is 94%. This indicates a very high fraction of assigned individuals have at least one ancestor born in the community enrichment region.

5 DISCUSSION

There are many factors which contribute to the success of a classifier. Sample size is one factor in building a strong classifier: In some cases the smallest communities had insufficient member examples to result in high precision and recall. In those cases, additional efforts may be needed to compensate for the small number of samples. However, there was no consistent trend between sample size and precision/recall, likely owing to the many other complexities related to the genetic, historical, and geographical separations between groups. For example, the small community of individuals with ancestors in New Mexico is in the 10th percentile for community size, but has a high F1 score which we attribute to its high degree of genetic separation in the IBD network. This classifier can obtain nearly 100% recall while maintaining over 90% precision. Another factor is genetic separation. For example, groups that have been separated by more subtle gene-flow barriers to surrounding populations, or who have many admixed individuals, such as the community of individuals with ancestors in Munster, Ireland, have lower F1 scores, but pedigree concordance scores above 90%.

Additionally, while the feature selection step may help protect against overfitting and reduces the computational complexity of building the classifiers, the number of features and method used to select features may influence the performance of the classifiers. Further exploration around the feature selection step and the impact of features on the classifiers themselves could prove insightful on the factors involved, and the behavior is likely to vary from community to community.

Several factors influence the pedigree concordance analysis. The first is community size; in smaller communities, there did not exist 100 individuals with pedigrees, limiting the data available. Further, the availability of historical records is non-uniform across sub-populations. Communities with fewer records have less of chance of having high concordance rates even with a perfect classifier. In some cases, the community is so geographically diffuse (e.g., a community of individuals with Jewish ancestry) that the concordance score will always be low and is not an appropriate measure of success.

While this work was successful at assigning individuals to multiple communities, the training dataset for the classifiers was limited to single-community assignments. While there have been subtle gene flow barriers across the United States in the last several hundred years, those barriers have likely decreased in strength due to Westward movement and improvement in modern transportation, resulting in many individuals in the United States with admixed ancestry (i.e., originating from many different ancestral populations). To support this, current population studies and census records in the United States show the ancestry of many individuals does not uniquely derive from a single sub-population [25]. A future area of investigation is the creation of a training dataset with individuals labelled as belonging to multiple communities to optimize community assignments for individuals with admixed ancestry.

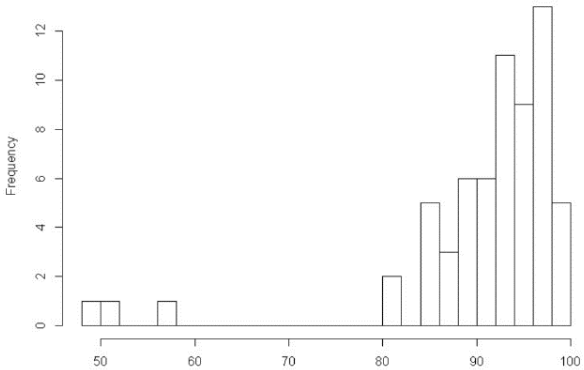

Figure 8. Pedigree Concordance Summary. The histogram shows the distribution of the pedigree concordance scores across all 63 community classifiers. Closer to 100 indicates that the ancestors of individuals classified as community members were born in the community enrichment regions.

Another area of future expansion will be in creating more diverse and balanced datasets. Since we embarked on this project, AncestryDNA's dataset has grown immensely and we are now able to obtain larger numbers of genotypes and pedigrees. The increase in genotypes and pedigrees may improve the granularity of community detection as well as improve the process of validation.

6 CONCLUSION

In this paper we have described a method that utilizes results from a large IBD network analysis to train scalable classifiers to assign individuals to the network's genetic communities, without needing to reconsider the entire network structure. A key challenge in network analysis is the constantly evolving nature of a network, and the processing required to discover structure in that network. In this paper, we show that carefully selecting features that describe the network structure can be used to train machine learning classifiers. This allows for a stable product and the ability to rapidly calculate assignments within that network. We are deploying a version of this system, and suggest that our approach may be broadly applicable in other network applications.

7 ACKNOWLEDGMENTS

This project has benefited from the contributions of the entire AncestryDNA Science Team, including Eunjung Han and Peter Carbonetto. The authors also wish to thank Keith Noto, Eurie Hong, Julie Granka, Oren Schaedel, Ariel Anderson, Michael Cormier, Harendra Guturu, Kristin Rand, Eyal Elyashiv, Yong Wang, Shiya Song, Jake Byrnes, Nathan Berkowitz, Benjamin Wilson, David Turissini, Daniel Garrigan, Shannon Hateley, Ladan Doroud, Natalie Myres, Cat Foo, Catherine Ball, and Kenneth Chahine for their contributions, helpful discussions, and feedback.

8 REFERENCES

[1] D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19:1655-1664, 2009.

[2] Ancestry Corporate Communications. Ancestry Sets AncestryDNA Sales Record Over Holiday Period and Fourth Quarter. *Press Release* available at: http://www.ancestry.com/corporate/newsroom/press-releases/ancestry-sets-ancestrydna-sales-record-over-holiday-period-and-fourth, 2017.

[3] C. Ball, *et al*. AncestryDNA Matching White Paper: Discovering genetic matches across a massive, expanding database. *Ancestry*. Available at: https://www.ancestry.com/corporate/sites/default/files/AncestryDNA-Matching-White-Paper.pdf

[4] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10(P10008), 2008.

[5] S. R. Browning and B. L. Browning. Haplotype Phasing: existing methods and new developments. *Nature Reviews Genetics* 12:703-714, 2011.

[6] C. Chen, A. Liaw, L. Breiman. Using Random Forest to Learn Imbalanced Data. *Statistics Technical Reports* 666, 2004.

[7] G. Csárdi and T. Nepusz. The Igraph Software Package for Complex Network Research. *InterJournal Complex Systems* 1695, 2006.

[8] G. Forman and M. Scholz. Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement. *SIGKDD Explorations*: 12(1), 2010.

[9] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:3-5:75-174, 2010.

[10] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12): 7821-7826, 2002.

[11] R. C. Griffiths and S. Tavare. The age of a mutation in a general coalescent tree. *Commun. Statist—Stochastic Models*, 14 (1&2), 273-295, 1998.

[12] A. Gusev *et al.* Whole population genome wide mapping of hidden relatedness. *Genome Research*, 2008.

[13] E. Han *et al.* Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications* 8, 2017.

[14] Illumina. Omni Whole-Genome DNA Analysis BeadChips. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_omni_whole-genome_beadchips.pdf, 2017.

[15] D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genetics* 8(e1002453), 2012.

[16] S. Leslie *et al*. The fine-scale genetic structure of the British population. *Nature* 519:309-314, 2015.

[17] B. K. Maples, S. Gravel, E. E. Kenny, and C. D. Bustamante. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *American Journal of Human Genetics* 93(2), 278-288, 2013.

[18] Moreno-Estrada *et al.* The Genetics of Mexico Recapitulates Native America Substructure and Affects Biomedical Traits. *Science* 344:1280-1285, 2014.

[19] M. Nei. Genetic Distance between populations. *Am. Nat.* 106: 283-292, 1972.

[20] M. E. Newman. The structure and function of complex networks. *SIAM Review* 45(2):167-256, 2003.

[21] R. Nielsen, J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, and E. Willerslev. Tracing the peopling of the world through genomics. *Nature* 541: 302-310, 2017.

[22] K. Noto *et al.* Underdog: A Fully-Supervised Phasing Algorithm that Learns from Hundreds of Thousands of Samples and Phases in Minutes. *Presented at the 64th Annual Meeting of the American Society of Human Genetics*, 2014.

[23] J. K. Pritchard, M. Stephens, P. J. Donnelly. Inference of population structure using multilocus genotype data. *Genetics* 155:945-959, 2013.

[24] J. S. Roberts *et al*. Direct-Consumer Genetic Testing: User Motivations, Decision Making, and Perceived Utility of Results. *Public Health Genomics*, 2017.

[25] US Census Bureau. 2010 Census Shows Multiple-Race Population Grew Faster Than Single-Race Population, https://www.census.gov/newsroom/releases/archives/race/cb12-182.html, 2012.