Inference and analysis of population structure using genetic data and network theory

Gili Greenbaum[1,2,*], Alan R. Templeton[3,4], Shirli Bar-David[2]

[1] *Department of Solar Energy and Environmental Physics, Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus 84990, Israel*
[2] *Mitrani Department of Desert Ecology, Blaustein Institutes for Desert Research, Ben-Gurion University of the Negev, Sede Boqer Campus 84990, Israel*
[3] *Department of Biology, Washington University, St. Louis, MO 63130, USA*
[4] *Department of Evolutionary and Environmental Ecology, University of Haifa, Haifa 31905, Israel*

* Corresponding author: gili.greenbaum@gmail.com

Abstract

Clustering individuals to subpopulations based on genetic data has become commonplace in many genetic and ecological studies. Most often, statistical inference of population structure is done by applying model-based approaches, such as Bayesian clustering, aided by visualization using distance-based approaches, such as PCA (Principle Component Analysis). While existing distance-based approaches suffer from lack of statistical rigor, model-based approaches entail assumptions of prior conditions such as that the subpopulations are at Hardy-Weinberg equilibria. Here we present a distance-based approach for inference of population structure using genetic data by defining population structure using network theory terminology and methods. A network is constructed from a pairwise genetic-distance matrix of all sampled individuals. The *community partition*, a partition of a network to dense subgraphs, is equated with population structure, a partition of the population to highly related groups. Community detection algorithms are used to partition the network into communities, interpreted as a partition of the population to subpopulations. The statistical significance of the structure can be estimated by using permutation tests to evaluate the significance of the partition's *modularity*, a network theory measure indicating the strength in which partitions divide the network. In order to further characterize population structure, a new measure of the Strength of Association (SA) for an individual to its assigned community is presented. The Strength of Association Distribution (SAD) of the communities is analyzed to provide additional population structure characteristics, such as the relative amount of gene flow experienced by the different subpopulations and identification of admixed individuals. Human genetic data are used to demonstrate the applicability of the analyses. The approach presented here provides a novel, computationally efficient, method for inference of population structure which does not assume an underlying model nor prior conditions, making inference potentially more robust. The method is implemented in the software `NetStruct`, available at https://github.com/GiliG/NetStruct.

1 Introduction

Inference and analysis of population structure from genetic data can be used to understand underlying evolutionary and demographic processes experienced by populations, and is therefore an important aspect in many genetic studies. Such inference is mainly done by clustering individuals into groups, often referred to as demes or subpopulations. Evaluation of population structure and gene flow levels between subpopulations allows inference of the migration patterns and their genetic consequences [1, 2]. As sequencing of larger portions of the genome is becoming more readily available, methods for such inference should ideally be able to take into account a large number of loci.

There are two types of approaches to clustering individuals based on genetic data: Model-based approaches and distance-based approaches. Model based approaches evaluate the likelihood of the observed data assuming that they are randomly drawn from a predefined model of the populations, for example that there are K sub populations and that these subpopulations are at Hardy-Weinberg equilibrium. Distance based approaches aim at identification of clusters by analysis of a matrix describing genetic distance between individuals or populations, for example by graphic visualization using multidimensional scaling (e.g. PCA), without prior assumptions. Over the last decade or so, model-based approaches have been more dominant as procedures for inference of population structure, mostly with implementation of Bayesian clustering techniques in programs such as `STRUCTURE`[3], `ADMIXTURE`[4] and `BAPS`[5]. It has been pointed out that distance-based methods have several disadvantages: they are not rigorous enough and rely on graphical visualization, they depend on the distance measure used, it is difficult to assess significance of the resulting clustering, and it is difficult to incorporate additional information such as geographical location of the samples[3]. Given these disadvantages, it would seem that distance-based measures are unsuitable for statistical inference of population structure. On the other hand, model-based approaches suffer from the necessity to restrict the interpretation of the results by heavily relying on the prior assumptions of the model, for example that the populations meet certain equilibria conditions, such as migration-drift or Hardy-Weinberg[3].

There has recently been a flourish of network theory applications to genetic questions in Genomics[6], landscape genetics[7] and migration-selection dynamics[8]. Recently, a network-based visualization tool (`NETVIEW`[9]) of fine-scale genetic populations structure, using a Super Paramagnetic Clustering algorithm[10], has been proposed and successfully applied to analysis of livestock breeds [11, 12]. However, this method still suffers from the many disadvantages of distance-based clustering approaches, and a more rigorous and statistically testable distance-based approach is still missing.

Development of a suitable distance-based network approach, that will not suffer from the disadvantages listed above, necessitates a clear definition of genetic population structure in equivalent network theory terminology. A genetically defined subpopulation is commonly thought of as a group of individuals within the population which are more genetically related (or more genetically similar) to each other than they are to individuals outside the subpopulation, as a result of many possible genetic processes such as migration, mutation and selection. In a network, a group of nodes which are more densely and strongly connected within the group than outside the group, relative to the given topology of the network, is called a ”community”. Therefore, in network theory terminology, the equivalent of a genetic population structure should be the community partition of a network constructed with individuals as nodes and edges defined using an appropriate genetic distance or relatedness measure. In network science, clustering nodes into groups has been extensively studied, and specifically community detection has attracted much interest[13]. Since the definition of a community is not rigid, and identifying optimal partitions is computationally expensive, many approaches and algorithms to optimally detect communities in networks have been proposed[14, 15].

We propose a network-based approach for analyzing population structure based on genetic data. We show that by applying recent advances in network theory, it is possible to design a distance-based approach that overcomes the previously described disadvantages of distance based approaches, and also does not suffer from the disadvantages of model-based approaches. We also show how rigorous statistical inference can be incorporated into this network-based approach in a manner that does not entail prior assumptions or conditions about the data. The process can be used with a large number of loci (e.g. microsattelites, SNPs) since it is computationally efficient in regards to the number of loci incorporated in the analysis. Moreover, we define a new measure for the strength to which an individual is associated with its assigned community, called Strength of Association (SA), and we show how Strength of Association Distribution (SAD) analysis can be used to infer further details regarding population structure, such as gene flow patterns of each subpopulation and identification of outlier individuals. The analysis is demonstrated on genetic data from human population extracted from the HapMap project[16]. In addition to presentation of a new distance-based alternative to population detection to be applied in population genetic studies, that complements existing model-based methods to give a more detailed and robust account of population structure, we believe that defining the problem of genetic population structure analysis in network terminology will allow future adoption and adaptation of network methods to address population genetic questions.

2 Methods

In this section we provide the relevant theory and describe a network-based approach for constructing genetic networks and inferring population structure by detecting community partitions on these networks. Following detection of community structure, we propose an additional exploratory analysis, based on a measure of the strength of association of individuals to communities, that may shed light on finer details of the community structure and therefore on population structure and underlying genetic processes.

2.1 Constructing networks from genetic data

A network can be described by an adjacency matrix, where the element in column i and row j is the weight of the edge connecting node i and node j. Therefore a genetic-distance matrix (a matrix describing some measure of genetic distance between all pairs of individuals, based on their genotypes) of a population can be regarded also as the adjacency matrix of a genetic-distance network. Many genetic distance and relatedness measures have been proposed[17], but if we restrict the discussion to symmetric relatedness measures (where relatedness between individual i and j is the same as between individual j and i), the genetic network thus described is a weighted undirected network (each edge is characterized by a weight but does not have directionality). Since we would like to extract information about the population structure from this network, we further restrict the discussion to genetic distance and relatedness measures which are expressed relative to allele frequencies in a reference population, i.e. measures that incorporate the allele frequencies of the total sampled population (local sampled populations allele frequencies should not be incorporated since this would mean that the null hypothesis is other than that there is no population structure).

In such a network, the strength of the connection between each dyad is relative to the genetic similarity between them, where shared rare alleles convey a stronger connection than do common alleles. Since even unrelated individuals may share many alleles, especially when many loci are examined, it is likely that this network will be extremely dense. It may therefore be useful, both from a computational point-of-view and in order to emphasize strong genetic relations within the population so as to increase detection power of network procedures, to remove edges which describe weak connections. This can be done in different ways, but the most straightforward approach is to remove edges with weights below a certain threshold, which is the approach we implement here. In this way a sparser network that consists of strong relatedness ties is attained.

Since using different thresholds will result in different networks which may give, for the analyses described below, different population structures, it is recommended to explore systematically different threshold values. For very low threshold values many weak relatedness ties will be included in the network, which may result in very dense networks which could mask densely related groups within the population. Very high thresholds may result in the network braking down into many disconnected components (a network component is a group of nodes that are connected within themselves but are not connected to any other node in the network), up to a point when the network includes only very small groups of connected nodes. Such networks are most likely not informative of population structure since they represented too few related dyads, and the community partition will likely consist of many one- or two-node communities (each community is confined to be within a component, and if the network consists of many small components then the community partition is constrained to include many small communities). Therefore the informative structures should be detected at the intermediate thresholds (see*Analysis of human SNP data* section for an example of a systematic exploration of threshold values).

2.2 Network communities and genetic population structure

In network theory, the term *community* refers to a subset of nodes in a network more densely connected to each other than to nodes outside the subset [18]. There are now several algorithms for efficiently partitioning a network into communities [14, 15]. Most commonly, a partition of a network into communities is evaluated by calculating the *modularity* of the partition, a quality measure (between -1 and 1) indicating whether the partition is more or less modular than would be expected if connections were randomly distributed[19]. The modularity of a particular community partition of a weighted network A is defined as the weight of the intra-community connections minus the expected weight of the intra-community connections in a random network preserving the edge weights of each node[20]:

$$Q = \frac{1}{A^*} \sum_{i,j} \left(A_{ij} - \frac{1}{A^*} \sum_k A_{ik} \sum_l A_{lj} \right) \delta(c_i, c_j) \tag{1}$$

where $A^* = \sum\limits_{k,l} A_{lk}$ is the sum over all edge weights in the network and $\delta(c_i, c_j)$ is a delta function with value one if nodes i and j are in the same community and zero otherwise. A positive modularity value indicates that the partition is more modular than expected. A partition of one community including all nodes results in a modularity of zero, and therefore for every network the optimal partition, maximizing the modularity, is always non-negative.

Since in a subdivided population the individuals in a subpopulation are expected to be more highly related in comparison to a random redistribution of relatedness levels between individuals, communities in the genetic network are expected to coincide with the subpopulations of the underlying population structure. We therefore propose that population structure can be ascertained by constructing a genetic network based on a genetic distance measure, and then applying one or several community detection methods to identify a partition which maximizes modularity. It is important to note that it is possible that the partition with the highest modularity is the entire network (with modularity of zero), and therefore community detection algorithms can also identify scenarios with no subdivision of the population.

Several approaches have been suggested in order to evaluate the statistical significance of community partitions [14]. However, since the genetic network, as described above, is constructed using multilocus genetic data, it is possible to pursue an alternative approach where the optimal modularity of the community partition can be compared with the modularity of community partitions of networks constructed from permutations of the genetic data. In this way it is possible to directly evaluate whether the modularity attained is significantly different than zero, and whether the network is significantly modular. This can be done either by permuting the genetic data (in each locus independently) and then constructing a genetic-distance matrix, or by permuting the genetic distance network while preserving the matrix symmetry (the latter is more computationally efficient when many loci are analyzed). Note that for a community partition of only one community, i.e. no population structure, the partition modularity is zero and therefore significance cannot be ascertained in this way.

2.3 Strength of Association Distribution (SAD) analysis

Revealing the division of the population into subpopulations may shed light on many aspects of the underlying evolutionary and ecological processes, however, more information can be attained by further analyzing the characteristics of the partition. The partitioning of the network into dense subgraphs, as presented above, does not convey information regarding how important each individual is to the detected partition. Here we introduce a measure intended to evaluate this aspect, the *Strength of Association* (SA) of individual i to its community. Given a community partition C and an individual i, we define the Strength of Association as

$$SA(C,i) = Q_C - \underset{\substack{k \\ C_k(i) \neq C}}{max}\ Q_{C_k(i)} \tag{2}$$

where Q_C is the modularity of the partition C and $C_k(i)$ is the partition identical to C except that node i is assigned to community k instead of its original community. Thus high SA values indicate that the partition C is sensitive to the assignment of i, and that the assignment of i to its community is essential, whereas low SA values indicate that there is another community that the individual is well assigned to. From a population genetic perspective, the measure evaluates how strongly individuals are related to the group to which they were assigned to, and SA is expected to be low when individuals are recent descendants from individuals from more than one subpopulation. Specifically, potential hybrids are expected to show low SA values, and the k that maximizes the term in equation 2 is the probable origin of the second lineage of the individual.

The SA measure is a measure at the individual level (although taking into account genetic data of the entire population). We introduce an exploratory subpopulation-level analysis that evaluates characteristics of subpopulations, the *Strength of Association Distribution* (SAD) analysis. This analysis examines the distribution of the SA values of the different communities and compares the statistical attributes of these distribution (e.g. the mean and variance of the SA values). Since different scenarios are expected to result in different cohesion of the subpopulations, it may be possible to infer what underlying processes where responsible for shaping the genetics of the population.

For example, a closed disconnected subpopulation is expected to display a narrow SAD with high mean (high community cohesion), since in a closed population individuals will be strongly related relative to the entire population, and individuals descended from lineages outside the subpopulation are rare. A subpopulation experiencing constant moderate gene flow levels is expected to display a wide or left-skewed SAD with high mean, since there should be many individuals with lineages that are mostly from the subpopulation, but recent migrant and descendants of recent migrants are expected to have low SA values, increasing the variance and the left-skewness of the distribution. A subpopulation experiencing constant strong gene flow levels is expected to display a wide SAD with low mean (low community cohesion), as many individuals will be descendants of migrants. A bimodal SAD distribution may indicate subgroups within the subpopulations experiencing different gene flow regimes as there are two groups corresponding to the two modes.

3 Analysis of human SNP data

In order to demonstrate the applicability of the network approach to infer population structure and the SAD analysis, we have selected a dataset of human SNP data, extracted from the Hapmap database [16]. This data set is well suited for the demonstration of a new approach since it is taken from a population where structure and demographic history are well known from archaeological, historic and genetic studies. The genetic data for this analysis consisted of 50 randomly selected individuals from each of the 11 groups in the HapMap project (overall 550 individuals). For each individual, 1000 polymorphic SNPs from each autosome were randomly selected (overall 22,000 sites per individual). In order to compare the results with a model-based approach, the same data were analyzed with the most widely-used model-based software,STRUCTURE[3].

3.1 Network construction

A genetic network was constructed from the genotypes (without any information on the original grouping of the individuals) using, for genetic distance calculation, a simple frequency-weighted allele-sharing relatedness measure. Analogous to the molecular similarity index[17, 21], we defined the frequency-weighted similarity at locus l for individual i with alleles a and b, with frequencies f_a and f_b (in the total sample) respectively, and individual j with alleles c and d:

$$S_{ij,l} = \frac{1}{4}\Big((1-f_a)(I_{ac}+I_{ad}) + (1-f_b)(I_{bc}+I_{bd})\Big) \tag{3}$$

where I_{ac} is one if alleles a and c are identical and zero otherwise, and the other indicators similarly defined. Note that this measure is commutative with respect to i and j. Given a sample with L loci, the weight of the edge connecting individuals i and j is defined as the mean frequency-weighted similarity over all loci:

$$G(i,j) = \frac{1}{L}\sum_{l=1}^{L} S_{ij,l} \tag{4}$$

The relatedness measure defined in equation 3 is a very simple symmetric relatedness measure, that measures diversity relative to the entire population, since it takes into account the allele frequencies at the level of the entire population (with sharing of rare alleles conveying a stronger connection than sharing common ones). Other, more sophisticated, measures are likely to construct more accurate networks and may be specific to the type of marker considered (e.g. for microsattelites the length of the repeat might be taken into account) or include additional information (e.g. geographic locations of the samples). The formulation presented here is designed to analyze diploid populations, but it can be easily generalized to any level of ploidy.

3.2 Community partition

There are currently many algorithms used for detecting population structure, relying on different network theory concepts (reviewed by Fortunato[14] and Lancichinetti[15]). We have used several of the commonly used algorithms (implemented using igraph[22]), presented in Appendix A, and we show here the results from the classic Girvan-Newman algorithm[13] (the results using different algorithms are not qualitatively different).

The edge-removal threshold parameter was systematically explored. For very low thresholds (0-0.181) the constructed networks were very dense and no population structure was detected (only one community was found which included all nodes). As mentioned above, this is to be expected from all but the most distinctly structured populations, since including connections between very weakly related individuals decreases the capability of the community detection algorithms to detect dense subgraphs within the network. For very high thresholds (above 0.209) the networks break down into many disconnected components, many of which include only one or two nodes. Such networks cannot be coherently analyzed for communities (see *Methods* section).

For the intermediate thresholds (0.182-0.208), different community partitions were detected for different threshold ranges. For thresholds in the range 0.182-0.188 two communities were detected, and Figure 1C shows results for threshold 0.188, referred to as "low threshold". For the range 0.189-0.195 three communities were detected, and Figure 1B shows results for threshold 0.194, referred to as "medium threshold". For thresholds above 0.196 the network was no longer connected and broke into several components, most notably a dense East Asian component and the rest of the network composed of one or more components. For the range 0.196-0.200 one community was detected in the East Asian component and four communities in the rest of the network. For thresholds above 0.201 only the East Asian component remained intact while the rest of the network broke into many small components and could no longer be meaningfully analyzed. The East Asian component consisted of one community for the threshold range 0.196-0.206 and two communities for the threshold range 0.207-0.208. Figure 1A shows the results of the community partition with threshold 0.207 of the East Asian component, with two communities, and for threshold 0.198 for the rest of the network, with four communities (referred to as "high threshold"). Within the ranges mentioned above there was no significant change in the assignment of the individuals to the communities. Therefore three qualitatively different community partitions of the network into communities have been found, with either two, three or six communities for low, medium and high thresholds respectively (Fig. 1).

Permutation test using 1000 permutations of the genotypes were conducted, and all community partitions were strongly significant ($p \leq 0.001$) for the three different community partitions (the test was not performed for the the partition of one community, see *Methods* section). With the low threshold the partition corresponded with an African\Non-African division (Fig. 1C), with the medium threshold to an African\Indo-European\East Asian division (Fig. 1B), and with the high threshold resulted in six communities: African, Indian, European, Mexican, Chinese and Japanese (Fig. 1A. Some of the other algorithms also detected the Masai population as a distinct community for the high threshold, Appendix A). The trend where higher thresholds reveals more detailed structure is correlated with the known broad patterns of human population differentiation. The low threshold coarse division of the population into two groups corresponds with the more ancient "out-of-Africa" migrations, the medium threshold additional division of the Eurasian population

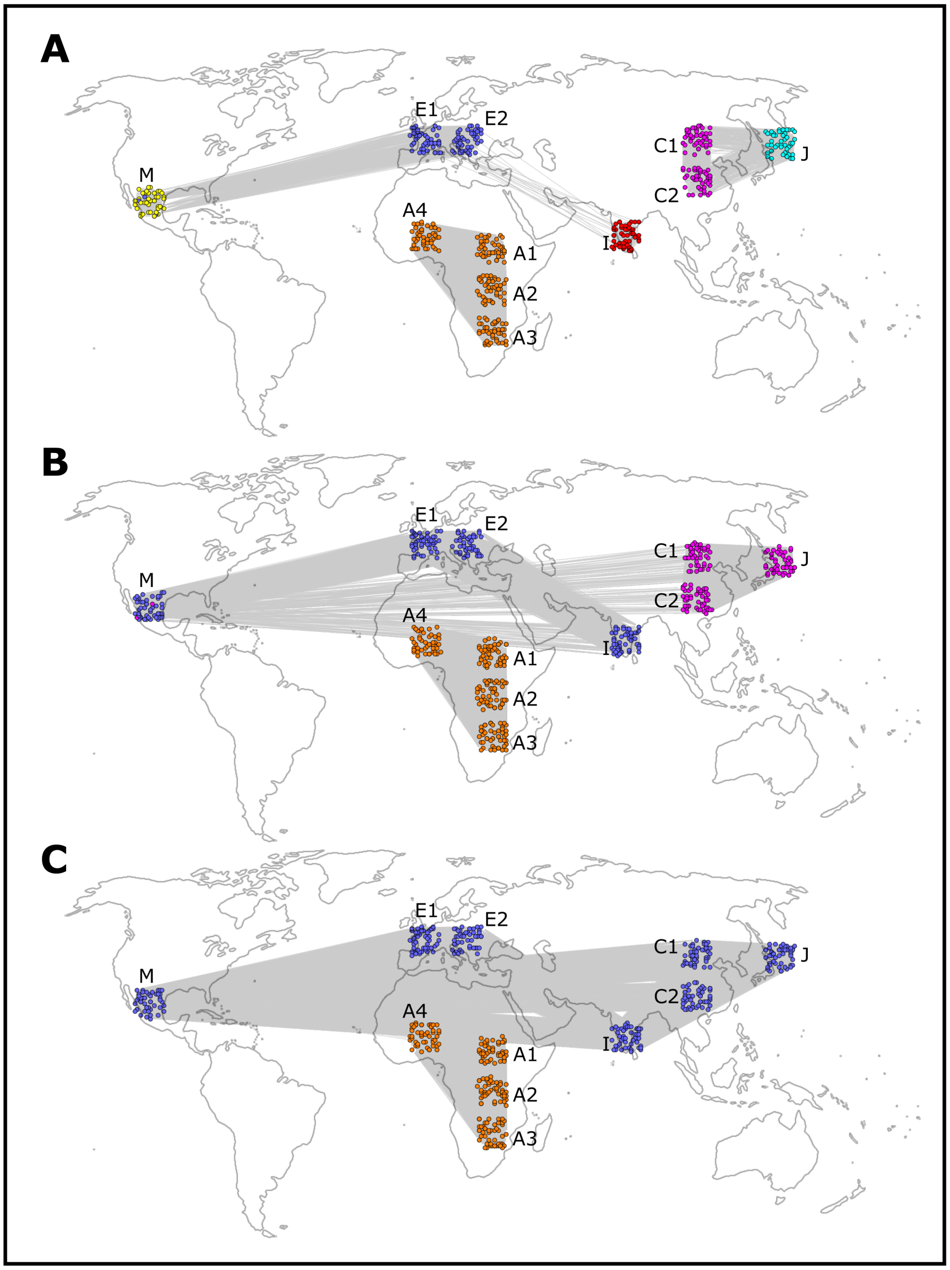

Figure 1: **Community detection on three networks with different thresholds.** Each node represents an individual, with colors representing the community assigned by the Girvan-Newman algorithm. (A) high threshold (0.207 for East Asian component, 0.198 for the rest of the network) (B) medium threshold (0.194) (C) low threshold (0.188). For visualization purpose, individuals are placed on the world map roughly corresponding to their ancestry. Sampled populations: A1 - African ancestry Americans; A2 - African (west Kenya); A3 - African (Masai); A4 - African (Nigeria); E1 - Europeans; E2 - European ancestry Americans; M - Mexican ancestry Americans; C1 - Han Chinese; C2 - Chinese ancestry Americans; J - Japanese; I - Indian ancestry Americans.

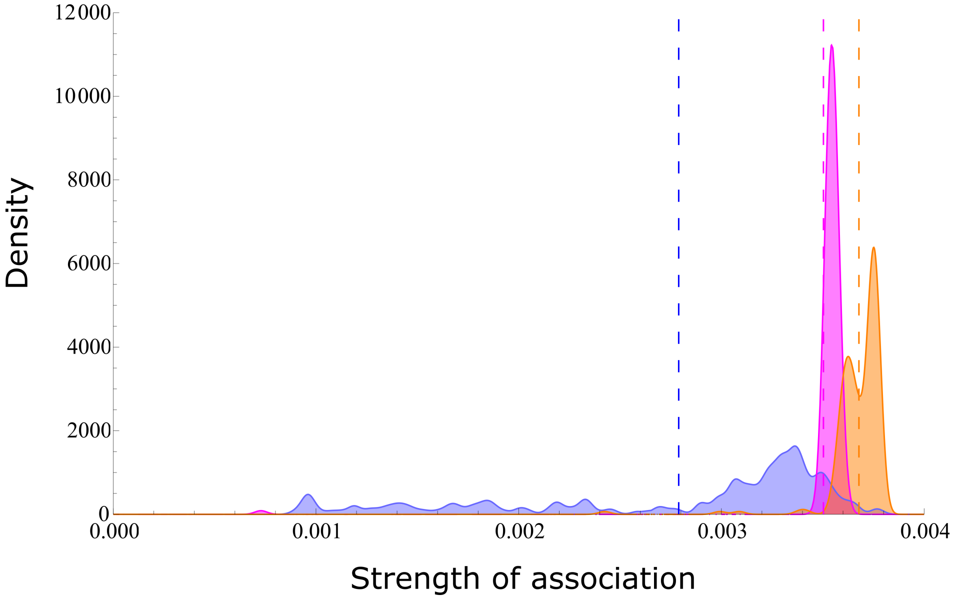

Figure 2: **SAD Analysis for the network in Figure 1B.** Shown are the distributions of the SA values for each of the three communities detected, with colors corresponding to the colors in figure 1B (orange to the African community, blue to the Indo-European community and Pink to the East Asian community). Mean SA indicated with dashed lines.

correspond with the more recent migrations to Asia, and lastly the high thresholds correspond with the most recent relevant migrations to India, Japan and Mexico.

The analysis with STRUCTUREwas done for different K values (K is the number of subpopulations assumed by the model). There is no statistical test available to evaluate the significance of the results for different models, but the most widely used heuristic is the one presented by Evanno[23]. This heuristic shows that the most likely K value is $K = 2$, but $K = 3$ and $K = 6$ are also indicated as likely values (Appendix B). For $K = 2$ and $K = 3$ the partition was the same as with the network approach. For $K = 6$ the detected partition consisted of five of the six subpopulations detected by the network approach: African, Indian, European, Mexican and East Asian. The Japanese\Chinese division was not detected, but the Masai individuals, assigned to the African subpopulations, were shown to be also likely to belong to a sixth subpopulation (results shown in Appendix B).

3.3 SAD analysis

As a demonstration of the SAD analysis, the network with medium threshold (Fig. 1B) was analyzed, and the distribution of the SA values for the three communities detected are shown in Figure 2. Figure 3 shows the equivilent analysis using the model-based results using STRUCTURE and assuming three subpopulations ($K = 3$).

The SAD of the East Asian community (pink) has a high mean and is a very narrow distribution, consistent with a subpopulation experiencing limited gene flow. This can be explained from the known historical trend of the relative isolation of East Asia from Europe and Africa.

The SAD of the Indo-European community (blue) is the one with the lowest mean SA, and is a wide left-skewed distribution, consistent with a subpopulation with defined core and periphery that experienced extensive gene flow relative to the other subpopulations. Given that the individuals belonging to this community are from European, Indian or Mexican ancestry, a probable explanation is that the core consists of the two European sampled populations and that the Indian and Mexican ancestry individuals have lower association with this group. This can be clearly observed when the distribution is decomposed to three distributions based on the sampled populations (Figure 4).

The distribution of the African community (orange) has a high mean, and is narrow and bimodal. This is consistent with a cohesive subpopulation with limited gene flow, but also that two distinct subgroups exist within the population with different levels of association to the community. Figure 5 shows the decomposition of the distribution to the four sampled populations composing it, and it can be seen that the the bi-modality can be explained by the fact that the Masai population (A3) is found to be a distinct population, as detected by some of the community detection algorithms (Appendix A). The STRUCTURE analysis also point out to this possibility, as for $K = 6$ it seems that the Masai individuals could possibility be assigned to a different subpopulation, although they are more likely to be assigned to the African subpopulation (Appendix B).

With the model-based analysis (Fig. 3), the African subpopulation is composed of the same individuals as in Figure 1B, however it can be seen that while individuals from A2 and A4 have almost no probability to be assigned to other subpopulations, individuals from A1 and A3 have non-negligable probability to be assigned to other subpopulations, which could be interpreted as that these two groups, while belonging to the African subpopulation, have experienced more gene flow from other subpopulations (mostly from the Indo-European subpopulation). The network analysis also finds that these groups are likely to have experienced more gene flow, as the mean SA for both these groups is lower than that of A2 and A4 (Figure 5). However, the distribution of A1 and A3 are quite different, which implies different evolutionary histories. The A1 SAD is skewed with a long left-tail, indicating that there are a number of individuals who are significantly less associated to the community and are possibly recently admixed individuals. The A3 SAD has a low mean but is symmetric without a tail, indicating that this population has experienced more gene flow but not in recent times. The recent admixture in A1, the African-ancestry Americans group, is consistent with recent higher gene flow experienced by the African ancestry Americans from other American groups.

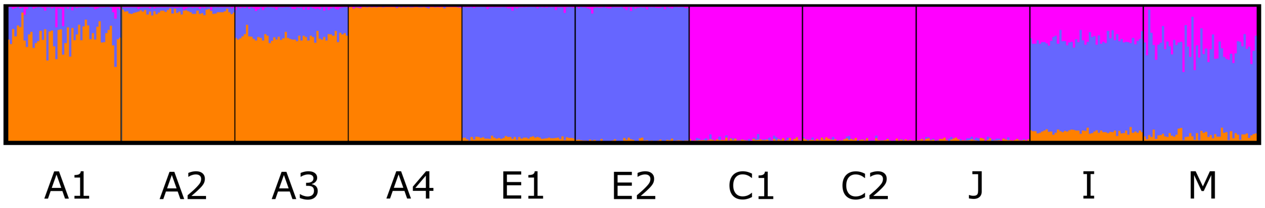

Figure 3: **Model-based analysis of human SNP data assuming three subpopulations ($K = 3$) using the program STRUCTURE.** The sampled population labels are the same as in Figure 1. The colors of the subpopulations correspond to the colors in Figures 1B and 2.

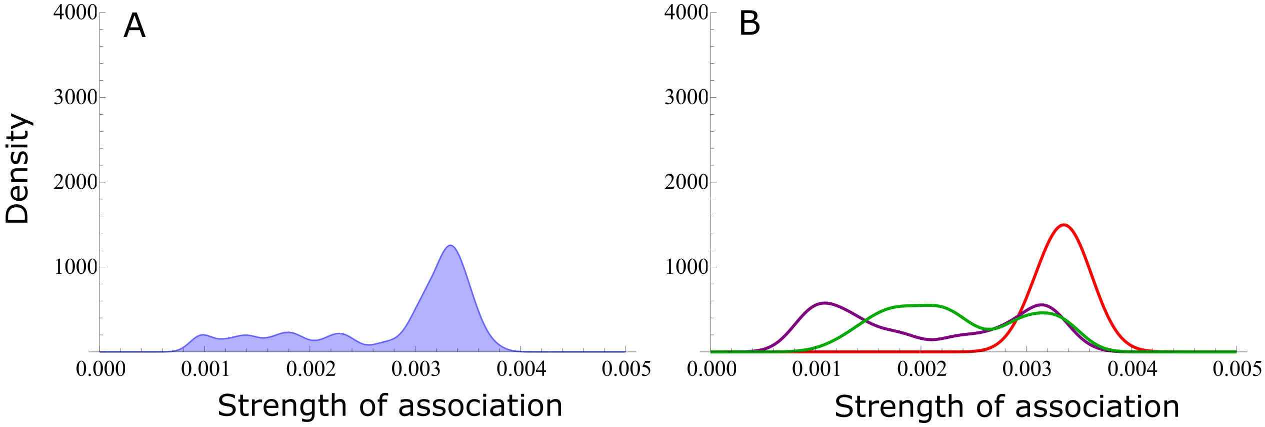

Figure 4: **SAD Analysis for the blue (Indo-European) community.** (A) Distribution of the blue community as shown in figure 2. (B) SA Distribution of the individuals in the community belonging to I (green), M (purple) and E1 or E2 (red). It can be seen that the individuals with European ancestry are responsible for the higher SA values in the distribution in (A), while the individuals with Mexican or Indian ancestry have lower association with this community.

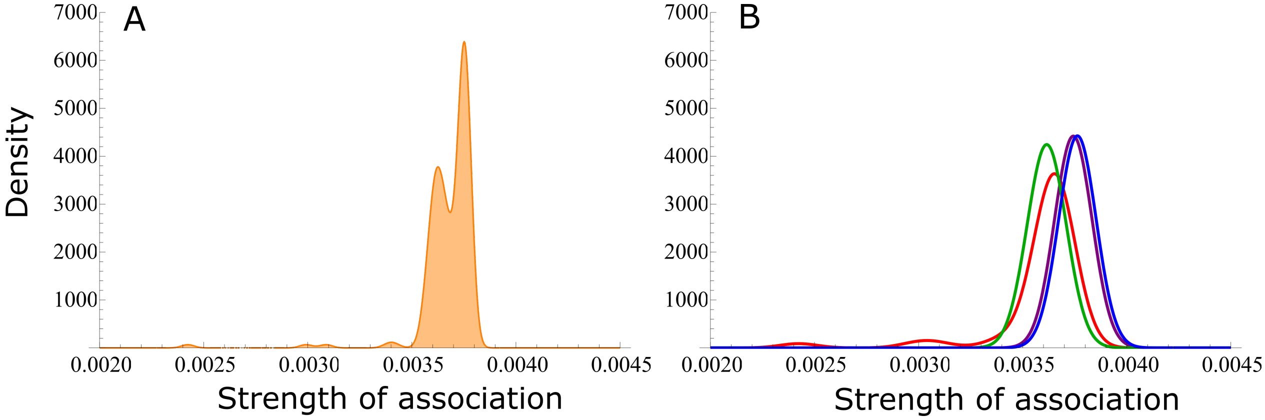

Figure 5: **SAD Analysis for the orange (African) community.** (A) Distribution of the orange community as shown in figure 2. (B) Distribution of the individuals in the community belonging to A1 (red), A2 (purple), A3 (green) and A4 (blue). The left mode in (A) is due to individuals from A3 (Masai) which was detected as a distinct population by some algorithms (appendix A). The individuals from America (red) have slightly lower association to the community than individuals from A2 and A4, as well as a distinct left-tail, probably because of recent admixture with people of European or Native American origin.

4 Discussion

We present a distance-based approach for analysis of population structure, which does not entail the assumptions of an underlying model or any prior conditions. The approach is set in a network theory framework and uses the concepts of *community* and *modularity*. The method allows computationally efficient assignment of individuals to sub populations, and is applicable also in cases where many loci are studied. An additional SAD analysis of the communities can be used to explore the population structure beyond assignment of individuals to populations by evaluating the strength in which individuals are associated with their assigned populations, which may be useful to detect admixed and outlier individuals as well as to explore finer details of the population structure. Potentially, inference of genetic and ecological processes from population structure detected by this approach should be more robust than inference from structure detected by model-based approaches, since no prior conditions are assumed. Ideally, population-level studies would benefit from exploring structure using our network approach in combination with Bayesian clustering methods and visualization by multidimensional scaling, as these complement each other and may give a more robust and detailed picture.In the example analyzed in this paper using human SNP data, the model-based analysis and the network analysis were relatively in agreement regarding assignment of individuals to subpopulations. However, the network analysis did detect the difference between the Chinese and Japanese groups which was not detected by the software STRUCTURE,and the SAD analysis revealed differences in gene flow experienced by the Masai and African-ancestry American groups that appear very similar in the STRUCTURE analysis (Figure 3).

One issue that has been a concern in model-based implementation is the assessment of the number of subpopualtions, K[3, 23], as K is usually one of the model parameters. By setting K these procedures regard the subpopulations as equivalent, even though this is often not the case. For example, for the network shown in Figure 1B, $K = 3$, however the three subpopulations show very different distributions of within-population relatedness (Fig. 2). In the network-based approach there is no such issue as most approaches for detecting community structure do not a priori assume K, but rather find the optimal K that maximizes modularity (e.g. [19, 24]), or acquires K as part of the detection process (e.g. [18, 25, 26]), without assuming any equivalence of the communities. Nevertheless, as has been demonstrated on human genetic data, using different thresholds for edge removal result in different K values (Fig.1), and the same is noticed when using different community detection algorithms (Appendix A). Since, in the analysis presented here, these are statistically significant community structures, this may reflect the fact that there is not necessarily a "correct" K value, but rather that different methods reveal structure at different hierarchical levels. Different significant community structures emerge, producing a semi-hierarchical structure, in the sense that a community partition at a given level does not depend on "higher" level partitions. True hierarchical community partition procedures[27, 28] can possibly be useful for hierarchical population structure analysis, but in most of these procedures each level is constrained by higher levels. It is important to note that the sampling scheme may also affect the number of subpopulations detected, as, for example, in a population with continuous isolation-by-distance gene flow pattern, sampling at discrete locations far enough apart may result in arbitrary K values which have no biological meaning[3].

With whole genome sequencing becoming more and more accessible, procedures for population structure analysis must also take into account computational considerations. The procedure presented here is composed of three consecutive steps, with construction of the network taking $O(Ln^2)$ time (n is the number of individuals in the sample and L is the number of loci involved in the analysis). Computation time for community detection depends on the algorithm used, but fast near-linear algorithms, taking $O(n + m)$ (m is the number of edges in the network) time and approaching $O(n)$ time for sparse networks (which can be constructed using low enough edge-removal thresholds), are already available[25]. SAD analysis depends on m and on the number of communities detected, c, and takes $O(cnm)$ time, approaching $O(cn^2)$ for sparse networks. Only the first stage involves the number of loci, therefore the computation time of the entire procedure is linear with respect to the number of loci, and there should be no computational limitation for including full genome sequences in analyses.

Since the genetic-distance measure, the threshold and the community detection algorithm remain, for now, used-defined, and may result in different population structures, care must be taken when defining these parameters, and preferably several options should be explored. Further studies may provide guidelines for setting these parameters as a function of the particulars of the system under study. Network theory, and particularly community detection, is a highly active field of research, but our understanding of the usefulness of particular community detection procedures to different types of networks is still minimal, and future advancements in network theory may provide clearer guidelines for algorithm and threshold choice.

We believe a network approach may provide an additional complementary viewpoint on population structure analysis, one less hampered by prior assumptions. Moreover, defining population genetic problems in network terminology is important in itself since currently many tools and methods are developed within the network theory framework in order to study complex systems. These methods may become accessible to the field of population genetics once network terminology is incorporated in population genetic theory and practice.

The method presented here is implemented in the program `NetStruct`, which uses community detection algorithms implemented in the software package igraph[22], and is available at https://github.com/GiliG/NetStruct.

References

[1] A. Templeton. *Population genetics and microevolutionary theory.* Hoboken, New Jersey: John Wiley & Sons, 2006.

[2] F. W. Allendorf, G. H. Luikart, and S. N. Aitken. *Conservation and the genetics of populations.* West Sussex: Wiley-Blackwell, 2012.

[3] J. K. Pritchard, M Stephens, and P Donnelly. “Inference of population structure using multilocus genotype data.” In: *Genetics* 155.2 (2000), pp. 945–959.

[4] D. H. Alexander, J. Novembre, and K. Lange. “Fast model-based estimation of ancestry in unrelated individuals”. In: *Genome Research* 19.9 (2009), pp. 1655–1664.

[5] J. Corander, P. Waldmann, and M. J. Sillanpää. “Bayesian analysis of genetic differentiation between populations”. In: *Genetics* 163 (2003), pp. 367–374.

[6] C. V. Forst. “Network genomics–a novel approach for the analysis of biological systems in the post-genomic era.” In: *Molecular biology reports* 29.3 (2002), pp. 265–280.

[7] R. J. Dyer and J. D. Nason. “Population Graphs: the graph theoretic shape of genetic structure.” In: *Molecular ecology* 13.7 (July 2004), pp. 1713–27.

[8] G. Greenbaum and N. H. Fefferman. “The potential applications of network methods to model selection-migration dynamics”. In: ().

[9] M. Neuditschko, M. S. Khatkar, and H. W. Raadsma. “NetView: a high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation.” In: *PloS one* 7.10 (Jan. 2012), e48375.

[10] M. Blatt, S. Wiseman, and E. Domany. "Superparamagnetic Clustering of Data". In: *Physical Review Letters* 76.18 (1996), pp. 3251–3254.

[11] A. Burren et al. "Fine-scale population structure analysis of seven local Swiss sheep breeds using genome-wide SNP data". In: *Animal Genetic Resources* 55 (2014), pp. 67–76.

[12] M. Neuditschko, M. S. Khatkarm, and H. W. Raadsma. "Fine scale population structure of global cattle breeds using dense haplotype data". In: *Proceedingsof the 10th World Congress of Genetics Applied to Livestock Production*. 2014.

[13] M. Girvan and M. E. J. Newman. "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12 (2002), pp. 7821–7826.

[14] S. Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3-5 (2010), pp. 75–174.

[15] A. Lancichinetti and S. Fortunato. *Community detection algorithms: A comparative analysis.* 2009.

[16] The International HapMap Consortium. "The International HapMap Project." In: *Nature* 426.6968 (2003), pp. 789–796.

[17] P. A. Oliehoek et al. "Estimating relatedness between individuals in general populations with a focus on their use in conservation programs". In: *Genetics* 173.1 (2006), pp. 483–496.

[18] M. Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Science* 103 (2006), pp. 8577–8582.

[19] M. E. J. Newman. "Detecting community structure in networks". In: *The European Physical Journal B* 38.2 (Mar. 2004), pp. 321–330.

[20] M. E. J. Newman. "Analysis of weighted networks". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 70.5 2 (2004), pp. 1–9.

[21] C. C. Li and D. G. Horvitz. "Some methods of estimating the inbreeding coefficient." In: *American journal of human genetics* 5.2 (1953), pp. 107–117.

[22] G Csárdi and T Nepusz. "The igraph software package for complex network research". In: *InterJournal Complex Systems* 1695 (2006), p. 1695.

[23] G. Evanno, S. Regnaut, and J. Goudet. "Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study". In: *Molecular Ecology* 14.8 (2005), pp. 2611–2620.

[24] A. Clauset, M. E. J. Newman, and C. Moore. "Finding community structure in very large networks". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 70.066111 (2004), pp. 1–6.

[25] U. N. Raghavan, R. Albert, and S. Kumara. "Near linear time algorithm to detect community structures in large-scale networks". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 76.3 (2007), pp. 1–11.

[26] P. Pons and M. Latapy. "Computing Communities in Large Networks Using Random Walks". In: *Computer and Information Sciences - ISCIS 2005*. Berlin, Germany: Springer, 2005, pp. 284–293.

[27] G. Palla et al. “Uncovering the overlapping community structure of complex networks in nature and society.” In: *Nature* 435.7043 (2005), pp. 814–8.

[28] R. Guimera et al. “Extracting the hierarchical organization”. In: *Proceedings of the National Academy of Science* 104.39 (2007), pp. 15224–15229.