Network-based hierarchical population structure analysis for large genomic datasets

Gili Greenbaum[1], Amir Rubin[2], Alan R. Templeton[3,4], and Noah A. Rosenberg[1]

[1] *Department of Biology, Stanford University,Stanford, California, United States of America*
[2] *Department of Computer Science, Ben-Gurion University of the Negev, Be'er-Sheva, Israel*
[3] *Department of Biology, Washington University, St. Louis, Missouri, United States of America*
[4] *Department of Evolutionary and Environmental Ecology, University of Haifa, Haifa, Israel*

Abstract

Analysis of population structure in natural populations using genetic data is a common practice in ecological and evolutionary studies. With large genomic datasets of populations now appearing more frequently, it is becoming increasingly possible to reveal many hierarchical levels of structure, including fine-scale genetic clusters. To analyze these datasets, methods need to be appropriately suited to the challenges of extracting multi-level structure from whole-genome data. Here, we present a network-based approach for constructing population structure representations from genetic data. The use of community detection algorithms from network theory generates a natural hierarchical perspective on the representation that the method produces. The method is computationally efficient, and it requires relatively few assumptions regarding the biological processes that underlie the data. We demonstrate the approach by analyzing population structure of the model plant species *Arabidopsis thaliana* from whole-genome data. This example illustrates how network-based approaches for population structure analysis are well-suited to extracting valuable ecological and evolutionary information in the era of large genomic datasets.

1 Introduction

Population structure is ubiquitous in natural populations, and the ability to detect it from information on genetic variation has been instrumental to our understanding of evolution and ecology. By studying population structure, researchers can investigate processes of gene flow and migration, quantify the evolutionary impacts of fragmentation and movement barriers, inform decisions concerning the conservation and management of populations, and interpret selection signals found in genomic studies. Currently, population structure analysis is an essential element in the toolkit of evolutionary biologists, conservation biologists, and ecologists, both for exploratory data analysis and for hypothesis testing [1–3].

As genome sequencing technology advances, large-scale whole-genome sequencing is becoming possible not only for describing genetic variation among species, but

also for studying differences and similarities among individuals within populations. Datasets describing genome-wide sequence variation for multiple conspecific individuals are now beginning to emerge [4–6], and a surge of whole-genome population datasets from wild populations is expected in the coming years. Whole-genome population structure analysis presents the promise of high-resolution population structure inference, such as multiple hierarchical levels and detection of fine-scale structure. However, the era of large genomic datasets also presents challenges for extracting and interpreting population structure at a genomic scale, as well as significant computational difficulties.

Network approaches have been introduced in recent years to model and analyze population structure [7–12]. In network-based population structure inference, the genetic relations between individuals within the populations are formulated as a network, incorporating the many complexities that result from evolutionary processes in natural populations, with few *a priori* assumptions regarding the biological processes involved in generating these relations. The resulting mathematical construct, the network, is then subjected to analyses developed in network science [13] to identify topological characteristics that can be interpreted as population structure. Network approaches have shown promise in the detail with which population structure can be described, as well as in computational efficiency [9].

Here, we describe a network-based approach for population structure inference that allows for efficient clustering of genetic information, producing hierarchical population structure diagrams, with application to large genomic datasets. Cluster representations containing many clusters at different levels in a hierarchy can then assist in evaluation and interpretation of the significance of the population structure detected. We provide an example analysis of population structure in a whole-genome dataset of *Arabidopsis thaliana*, emphasizing the applicability of our approach for studying ecology and evolution.

2 Results

2.1 Detecting hierarchical population structure using networks

The problem of detecting population structure from genetic data has a particular emphasis on the hierarchical nature of many natural population structures. In order to use networks to detect population structure, we first formulate genotype data as a genetic similarity network, with individuals as nodes and inter-individual genetic similarities as edges. To construct a genetic similarity network, we use a frequency-weighted allele-sharing similarity measure [9], with the weighting scheme designed

to increase the contribution assigned to shared rare alleles. Constructing pairwise genetic similarity networks for large numbers of loci is a computationally intensive process; however, it is also highly parallelizable and can therefore be completed in reasonable time with the readily-available computation clusters available today, even for whole-genome datasets (see *Computational efficiency* section in *Methods*).

In our approach, the problem of detecting population structure — groups of individuals sharing common evolutionary histories — is equated with the problem of detecting dense subnetworks within the genetic similarity network. In other words, we identify groups of individuals that are highly interconnected. In network science, such dense groups are termed "communities," and many community-detection methods have been developed in recent years [14–16].

The genetic similarity network summarizes genetic variation from across the entire genome. However, the network is expected to be extremely dense, with most pairs connected, because any two individuals, even if belonging to distinct subpopulations, are expected to share some alleles inherited from distant common ancestors. This property adds much noise to any signal of structure, potentially decreasing the ability to detect fine-scale structure. In particular, fine-scale population structure is characterized by smaller groups of individuals that share a more recent genetic history within the group than do larger groups representing coarser-scale structure. Therefore, genetic similarity within fine-scale clusters, which are expected to contain only strong genetic similarities, is expected to be higher than that in coarser clusters, which are expected to contain weaker but nontrivial network connections. Consequently, we explore population structure across many hierarchical scales by iteratively pruning weak edges from the network and applying community detection procedures (Fig. 1). By removing weak edges from a coarse cluster, the population structure signals characteristic of fine-scale clusters within the coarser cluster become distinguishable, and another hierarchical level can be revealed. The resulting set of clusters can be described as a tree-like hierarchy of nested clusters, which we term a "Population Structure Tree" (PST). In a PST, the root cluster contains all individuals, the leaf clusters describe the finest-scale structure, and the clusters represented by internal nodes describe intermediate levels of clustering (Fig. 1). The internal clusters overlap with clusters emerging from them as "descendants", while all the leaf clusters are non-overlapping. Unlike species trees, which usually describe independent evolutionary histories along the branches, in PSTs, no assumptions are made regarding the evolutionary processes that give rise to the hierarchical structure, and the branches need not represent groups that have evolved in isolation.

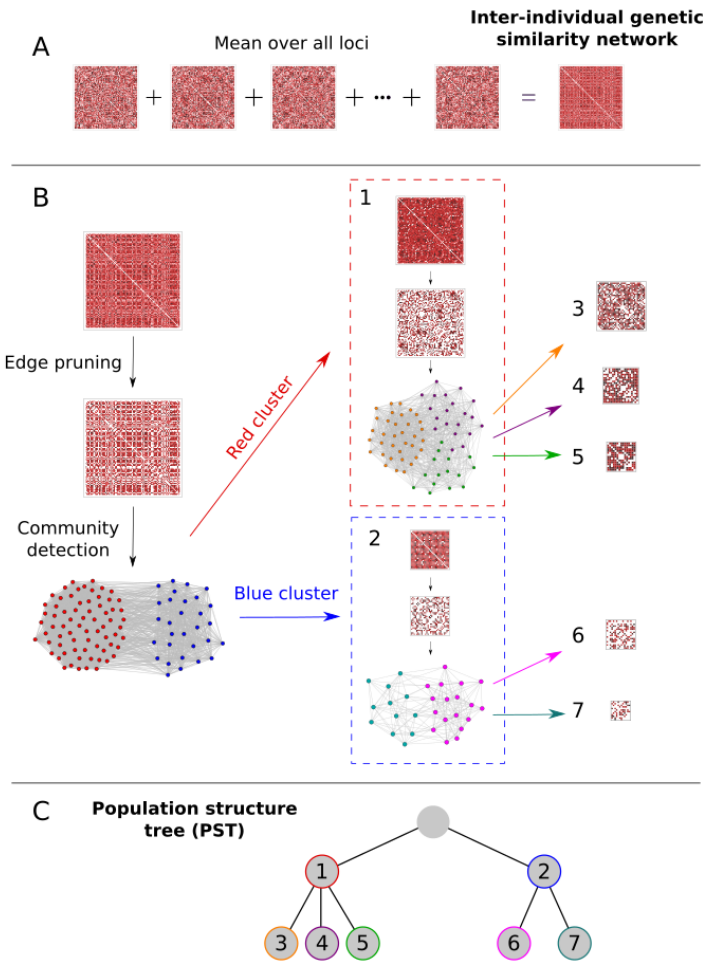

Figure 1: Schematic representation of a network-based construction of a population structure tree (PST) from genomic data. (A) For each SNP, an inter-individual genetic similarity network (adjacency matrix) is constructed using a frequency-weighted allele-sharing genetic similarity measure (Eq. 1). The mean over all loci is taken, to produce a genome-wide genetic similarity matrix. (B) Weak edges are pruned from the matrix, by setting low matrix entries to 0 until a community structure emerges, as detected using network community-detection algorithms. Each community (numbered submatrices) is then analyzed independently in a similar manner. Notice that finer-scale clusters are characterized by darker matrices, indicating structures characterized by higher genetic similarities. (C) The analysis is summarized as a PST diagram, summarizing the hierarchical levels of population structure and their relationships.

2.2 Visualization

An important aspect of interpreting population structure analyses is visualization of the outputs. We visualize population structure on geographic maps using a coloring scheme designed to reflect the topology of the inferred PST. We start with a color interval (e.g. "rainbow colors"), and assign it to one of the clusters, which becomes the root cluster for that coloring. Our coloring scheme seeks to describe the topology of the branch connecting the selected root cluster and the leaves. Each daughter cluster (a cluster emerging from a cluster through the edge-pruning and community detection process) of the selected root cluster inherits an equal fraction of the color interval, proportional to the number of daughter clusters, in no particular order. This process of transmission of color intervals to daughter clusters is then repeated, until the leaves of the hierarchy are reached on all branches. Each cluster is then assigned a color that corresponds to the midpoint of its assigned color interval. This scheme results in a coloring of a branch of the PST, such that closer clusters in the hierarchy, within and between hierarchical levels, are assigned closer colors on the initial color interval (see examples in Fig. 2 and Fig. 3).

Once the clusters are assigned colors, each individual is assigned a color corresponding to the finest-scale cluster to which it is assigned. On a geographic map of the region of interest, each individual is placed in its geographic sampling position, and colored with its assigned color. This step results in a map reflecting the topology of a particular branch of the PST. By selecting different root clusters and repeating the process, different geographic regions and different levels of structure can be emphasized (e.g., Fig. 2 and Fig. 3).

2.3 Comparing population structure trees

Large genomic datasets provide opportunities to conduct subsampling analyses on the same dataset, for example to investigate population structure inferred from different genomic regions, or for studying the effect of different genome sequencing schemes. To interpret such subsampling analyses, a coherent method for comparing population structure outputs is needed. We therefore developed an approach to compare different PSTs derived from the same dataset.

The approach is based on normalized mutual information (NMI), an information-theoretic measure that evaluates, for a pair of variables, the amount of information gained about a variable by observing a second variable. This measure can be adjusted to apply to a set of hierarchical sets of clusters [17], such as PSTs. Given two PSTs derived from the same set of individuals but not necessarily from the same similarity matrix, the measure produces a score that ranges from 0 to 1, where higher scores indicate greater similarity between the PSTs (see *Methods* for details). Notably, two

PSTs might differ both in the topological ordering of the clusters in the tree and in the assignment of individuals to clusters, but they are still comparable using the NMI measure.

Our implementation of the NMI measure is flexible in the sense that it is possible to conduct comparisons not only between two PSTs, but also between subsets of the PSTs. This type of comparison enables a focus on comparing certain features of the PSTs. For example, the approach can compare specific branches of interest, or only fine-scale leaf clusters, by considering in the NMI comparisons specified subgroups of clusters (see *Methods* and Supplementary Information).

2.4 Population structure tree of *Arabidopsis thaliana*

Understanding the population structure of the model plant species *A. thaliana* has been of great interest both for its value for explaining patterns of natural variation and for correct interpretation of signatures of selection and association inferred from genome-wide association studies [5, 18–26]. Within Eurasia, population structure has been characterized in one study by nine clusters — the group of highly differentiated individuals found primarily in the Iberian Peninsula, and eight other clusters, broadly corresponding to North Sweden, South Sweden, the Iberian Peninsula, Asia, Germany, Italy/Balkans/Caucasus, West Europe, and Central Europe [5]. Recently, evidence provided by samples from Africa has suggested that African populations are distinct from Eurasian populations [26]. In North America, the population has been found to be relatively unstructured [5, 19, 27].

Here, we analyzed worldwide population structure of *A. thaliana* using our network-based method, in order to better understand hierarchical levels and fine-scale structure in this model species. We studied 1,135 individuals (accessions) from Eurasia and North America [5], and 79 individuals from Africa [26], with the combined dataset containing 1,214 individuals and 2,367,560 SNPs.

We detected a highly hierarchical population structure, with 247 (overlapping) clusters overall and 96 fine-scale (non-overlapping) leaf clusters at the tips of the hierarchy (Fig. 2A). The topology of the inferred PST is highly unbalanced, with one main branch consisting of a single cluster (left main branch in Fig. 2A), and the other main branch containing 246 clusters (right main branch in Fig. 2A). All individuals in the single-cluster left main branch are from North America, representing 71% of the North American sampled population (89 of 125). The right main branch consists of two sub-branches, one representing European individuals (Fig. 2B) and the other representing Asian, African, and North Swedish individuals, and some Iberian individuals designated as "relicts" by previous studies [5, 25, 26] (Fig. 2C). Population structure in the primary European branch shows relatively continuous genetic differentiation over space, as evidenced by the fact that geographically close samples

are closely positioned in the PST topology (Fig. 2B). In the primarily non-European branch, population structure is more discontinuous geographically, with a large "gap" between the Iberian, North Swedish, and Asian branches (Fig. 2C). However, looking at fine-scale population structure in Africa and Iberia (Fig. 3F-3G), and in Asia (Fig. S1), population structure appears to be more geographically continuous within these smaller branches. Note that in North Sweden, the sampled area was too small to examine geographic patterns of differentiation within the region.

Structure can also be ascertained at a very fine scale by examining branches close to the tips of the hierarchy. In the Iberian Peninsula, for example, genetic differentiation can be observed at several scales (Fig. 3). At the broader scale, in the western part of the peninsula, genetic differentiation is aligned along a north-south gradient (Fig. 3B), whereas in the northeastern region, genetic differentiation follows more of an east-west gradient (Fig. 3C). Distinct groups, correlated with geography, can be seen at fine spatial scales, such as along the northeastern coast (Fig. 3E), in the north (Fig. 3B), and in the central region (Fig. 3D). In Morocco and the Iberian Peninsula, for those individuals belonging to the primary non-European branch in Figure 2A, population structure follows a north-south gradient, with clusters corresponding to small and specific regions in Morocco (Fig. 3G) and to larger regions in the Iberian Peninsula (Fig. 3F). In Asia, population structure corresponds to broad regions (e.g. Ural Mountains, Altai Mountains, Central Asia), but finer-scale patterns are also observed within these regions (Fig. S1).

Although the single cluster associated with North America represents most of that population, 29% of the North American samples are clustered within the European branch: 20 in a UK/France cluster, 14 in a Central European cluster, and 2 in a South-Sweden/North-Europe cluster (Fig. S2). These samples include 5 of the 6 individuals sampled along the western coast of North America.

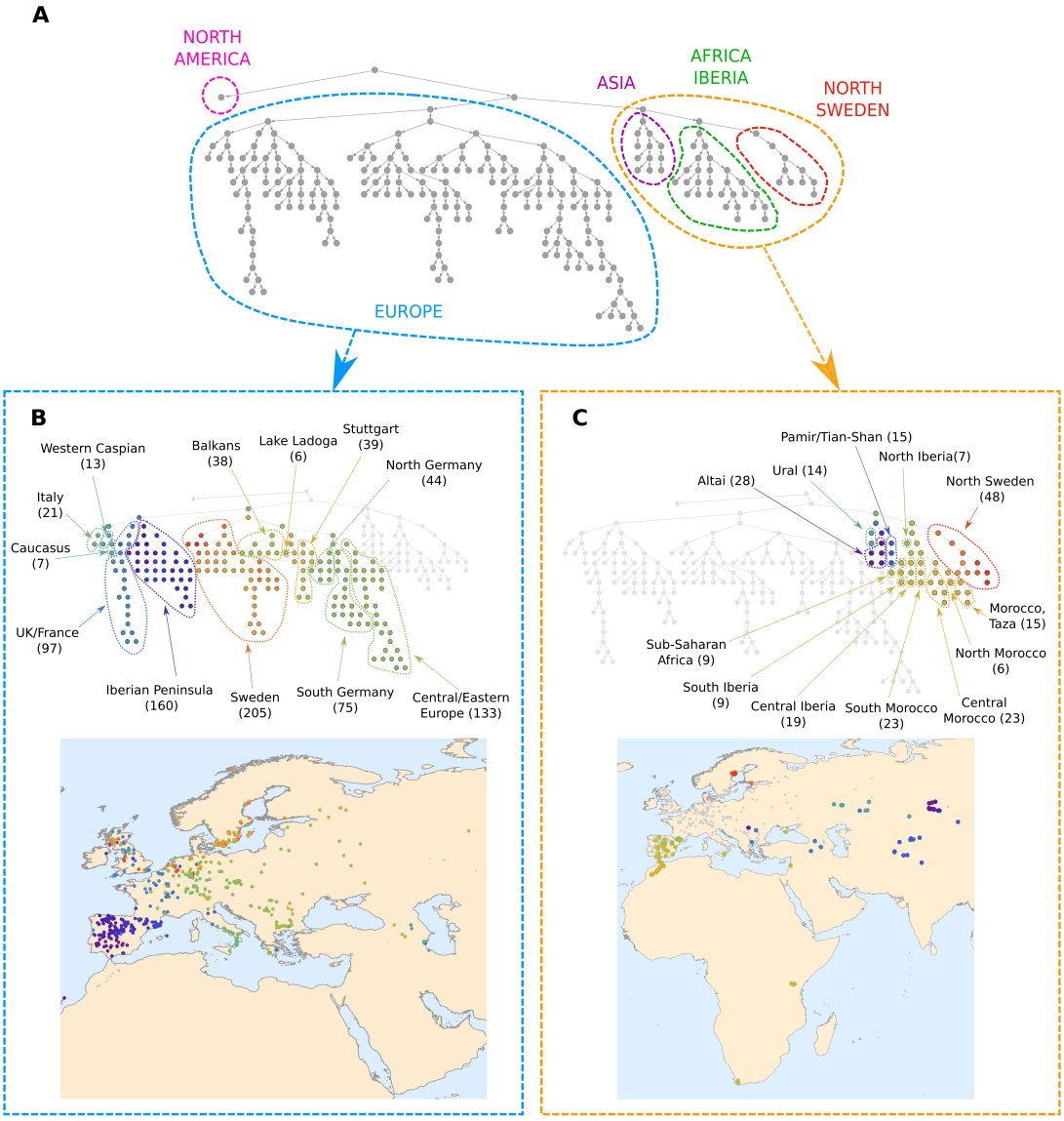

Figure 2: Broad-scale population structure of *A. thaliana* from network-based whole-genome analysis. (A) The inferred population structure tree (PST). Each element in the hierarchy represents a cluster of individuals, and each cluster contains those clusters below it in the hierarchy. The root element represents the entire sample of 1,214 individuals. In colored dashed lines, the main regions corresponding to sampling locations are indicated (with the labels defined *post hoc*). (B) Visualization of the branch corresponding to most European sampling locations. Each cluster is assigned a color such that "closer" colors represent closer clusters in the PST. On the map, each individual is placed at its sampling location and colored according to the finest-scale cluster to which it was assigned. (C) Visualization of the branch corresponding to Asia, Africa, North Sweden, and some samples in the Iberian Peninsula.

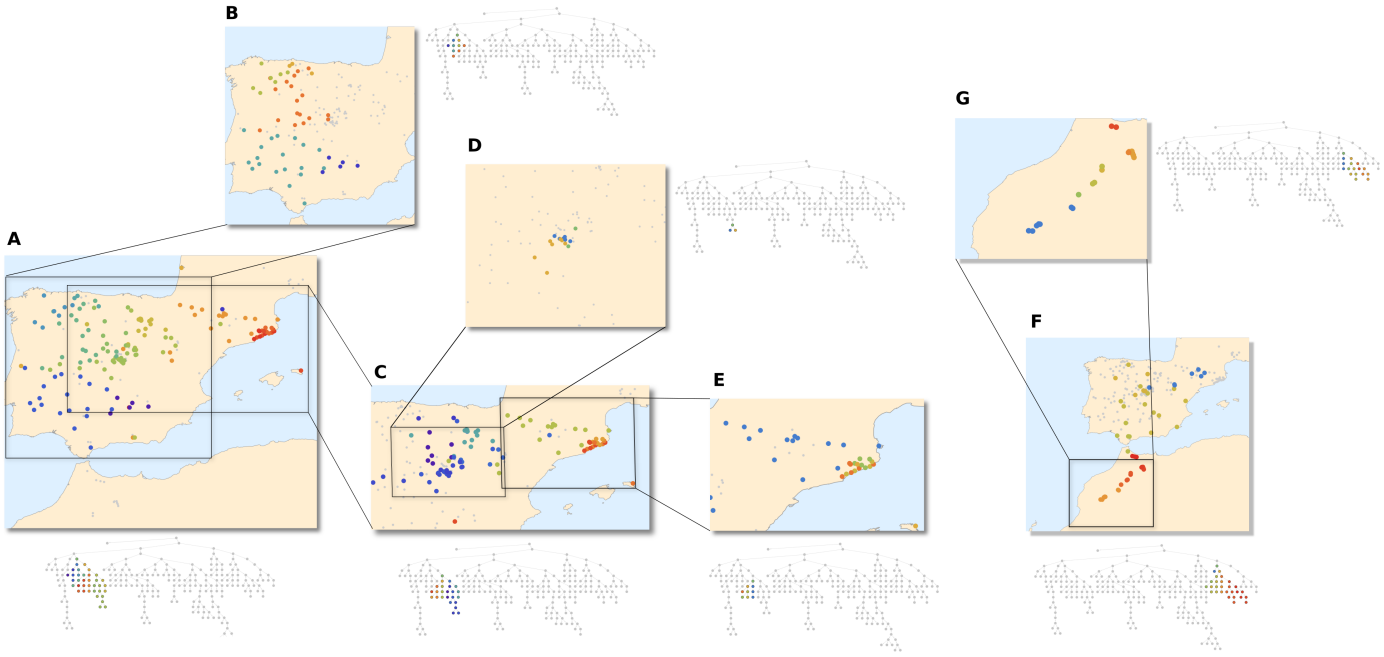

Figure 3: Fine-scale population structure of *A. thaliana* in the Iberian Peninsula and Morocco. On each map, a branch of the inferred PST from Figure 2 is visualized. Adjacent to each map is the PST colored in the same manner as in the map. Panels (A)–(E) show a sub-branch of the primary European branch (blue branch in Figure 2A), and panels (F) and (G) show a sub-branch of the primarily non-European branch (orange branch in Figure 2A). (A) Branch corresponding to most of the Iberian population (Iberian peninsula branch in Fig. 2B). (B) Branch corresponding to the western part of the Iberian peninsula, with differentiation along a north-south gradient. (C) Branch corresponding to the northeastern part of the peninsula, with differentiation along an east-west gradient. (D) Branch corresponding to a small area in the center of the peninsula. (E) Branch corresponding to north Spain, with a differentiated sub-branch along the northeast coast. (F) Branch not belonging to the primary European branch in Figure 2A, corresponding to the Iberian Peninsula and Morocco. Population structure extends on a north-south axis from Africa to Europe over the Strait of Gibraltar. (G) Branch corresponding to Morocco, showing fine-scale population structure in Morocco.

2.5 Information gain by increased SNP coverage

We estimated the effect of increasing the number of loci used in the inference of PSTs by comparing the PST inferred from whole-genome data with PSTs inferred from subsamples of the genome, using the information-theoretic NMI measure. This step was accomplished by evaluating NMI between PSTs, one inferred from the full dataset as in Figure 2, and one inferred from a subsample of the data. We also computed NMI for fine-scale structure by observing only the leaves of the hierarchies. For this purpose, without considering SNP positions, we randomly subsamped all SNPs to produce subsets of 0% to 98% of the original genome in increments of 2%. For each frequency value, 100 random subsamples were taken, and from each of these subsamples a PST was inferred and compared to the full-data PST.

The results (Fig. 4) indicate that information gain is rapid for the first 10% of the genome length, and further SNP coverage results in only modest information gain. The sharp decrease in information gain above 10% SNP coverage suggests that most population structure is revealed in this dataset using about 250,000 SNPs. These results are consistent with previous studies on lower-density genetic data that have shown that the resolution of population structure detected tends to saturate when enough loci are included in the analysis [28–33]. Beyond 10% SNP coverage, more information is gained regarding fine-scale structure than coarse-scale structure; for example, increasing the number of SNPS from 236,756 (10% SNP coverage) to 2,130,804 (90% SNP coverage) increases NMI by only 3.7% when considering the entire PST, whereas the increase is larger when considering the finest-scale clustering at the PST leaves, with 11.1% increase in NMI (Fig. 4).

3 Discussion

3.1 Population structure inference using networks

The evolutionary processes shaping genetic variation are often complex, occurring at multiple spatial levels and over multiple time periods. Network-based methods are well-suited for population structure inference since they enable a natural hierarchical perspective. Our approach is data-driven and does not require assumptions or pre-specified models regarding the biological processes responsible for shaping the genetic patterns. The visualization technique presented here, with PST diagrams and corresponding map colorings, is useful for summarizing information on population structure while preserving information on both broad and fine scales.

We also demonstrated the applicability of the NMI measure for comparing PSTs to understand the information gained by an increased number of genetic markers. This measure can also be useful for other types of analyses, such as for comparing

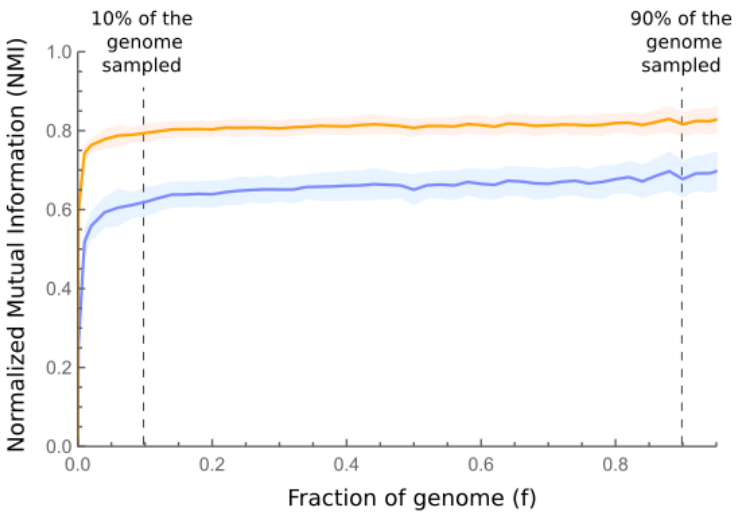

Figure 4: Normalized mutual information (NMI) between the PST inferred using the entire genome and PSTs inferred from a subsampled fraction of the genome. The NMI values evaluate the amount of information gained on hierarchical population structure by sampling a fraction of the genome, f, relative to sampling the entire genome. The mean NMI across 100 random subsamples for each f value considering the entire PST topology is shown in orange; mean NMI considering only the finest-scale leaf clusters is shown in blue. Shaded regions show standard deviations across the 100 replicates. NMI values saturate at values below 1 because cluster assignments often switch at fine scales (e.g. between PST leaves) when PSTs are inferred from subsampled data.

PSTs inferred from different genomic regions or different types of genetic markers.

There has been a long history of interest in both model-based and data-driven, non-model-based, approaches for population structure inference, with advantages reported for each of the two classes of methods [9, 34–37]. Our network approach falls into the latter category, and it is therefore desirable to be used in combination with model-based approaches that provide a different and complementary perspective on population structure representation.

3.2 *A. thaliana*

Our analysis of the *A. thaliana* dataset presents a complex picture of population structure across Africa and Eurasia, with relatively little population structure in North America. Particularly interesting is the separation, at a high hierarchical level, between the European branch and the African, Asian, and North Swedish branch (Fig. 2). Such a separation could reflect an ancestral split between these groups [26], but how these spatially distant populations are related, and why a geographic gap ex-

ists in the primarily non-European branch, remain to be investigated. Further study of the population in North Sweden, and increased geographic sampling in Africa and Asia, could help expain these patterns. The continuous differentiation, across hierarchical scales, in the European branch accords with past studies of this geographic region [5, 18, 20, 22].

In contrast to the multi-level and fine-grained population structure detected in Eurasia, most of the population in North America is assigned to a single cluster (Fig. 2A). This lack of structure, which acords with previous studies [19, 23, 27], could result from a founder effect followed by rapid expansion, producing a fairly homogeneous but widespread population [23]. However, a significant portion of the North American population (29%) was not assigned with the other North American individuals, suggesting a history more complex than a single colonization event. Most individuals perhaps descend from a founder population in the eastern United States [18, 19], but other migration events, possibly from Western European sources, might also have contributed to the population.

Fine-scale structure was detected in many regions, at local spatial scales. The formation of such small-scale structure might be attributable to small-scale landscape barriers and resistance to gene flow, but also to adaptation of *A. thaliana* to specific local conditions [38–40]. For example, the branch corresponding to the coastal population in northeastern Spain has been suggested to be locally adapted to high salinity [41]. Selective sweeps suggesting significant local adaptation have also been reported for the population in North Sweden [42, 43]. Such events could have led to decreased fitness of hybrids and decreases in effective migration, intensifying the differentiation of this population. This phenomenon could perhaps contribute to explaining the non-intuitive placement of North Sweden in the PST, separate from most European samples (Fig. 2C).

3.3 Conclusions

We have presented a network-based approach for revealing population structure at multiple hierarchical levels, summarized as a PST, visualized on geographic maps, and applicable to whole-genome datasets. The approach is data-driven and computationally efficient, allowing analysis of large genomic datasets. Our method for comparing PSTs is potentially useful for conducting studies that investigate population structure inferred from subsamples of the data.

4 Acknowledgments

We thank Arun Durvasula, Angela Hancock, and Andrea Fulgione for help with the *A. thaliana* data. We also thank Ellie E. Armstrong, Stefan Prost, and Danny Hendler for helpful discussions. This study was supported by NIH grant R01HG005855 awarded to NAR.

5 Author contributions

GG conceived the study and drafted the manuscript. GG and AR developed and implemented the approach, and analyzed the *A. thaliana* data. ART and NAR contributed ideas to the framing and implementation of the project. All authors read and commented on the manuscript.

6 Data availability

The code for running the analyses (called NetStruct_Hierarchy) can be found in `https://github.com/amirubin87/NetStruct_Hierarchy` along with a user manual and a tool for plotting maps and coloring the clusters. The *A. thaliana* genetic data was extracted from `www.mpipz.mpg.de/hancock/downloads`, which stores a unified datast of the 1001 Genomes database (also found at `1001genomes.org`) and the data from Durvasula et al. [26]. Geographic locations of the samples were extracted from the 1001 Genomes database at `1001genomes.org` and from Durvasula et al. [26].

7 Methods

7.1 Constructing genetic-similarity networks

The first step of network-based population structure detection is to use genomic data to generate a genetic similarity network, with individuals treated as nodes and inter-individual genetic similarities as edges (Fig. 1). Here, we use a frequency-weighted allele-sharing similarity measure. Allele-sharing measures count the number of shared alleles between each pair of individuals over all loci. Because individuals that share rare alleles are more likely to belong to the same subpopulation than are those that share common alleles, we assign more weight to shared rare alleles than to shared common alleles.

Considering a locus ℓ, for individual i with alleles a and b and individual j with alleles c and d, the frequency-weighted allele-sharing similarity is defined as [9]:

$$A_{ij,\ell} = \frac{1}{4}\left[(1-f_a)(I_{ac}+I_{ad}) + (1-f_b)(I_{bc}+I_{bd})\right], \tag{1}$$

where $I_{xy} = 1$ if x and y are identical and $I_{xy} = 0$ otherwise, and f_a and f_b are the frequencies of alleles a and b in the entire sample. The genetic-similarity edge connecting individuals i and j is weighted by the mean genetic similarity over all loci available for this pair of individuals, $A_{ij} = \frac{1}{\ell}\sum_{\ell} A_{ij,\ell}$. When some data are missing, A_{ij} is computed based on all those loci for which both individuals i and j are scored. This measure is symmetric ($A_{ij} = A_{ji}$) and defines an inter-individual genetic-similarity network, where A_{ij} is the adjacency matrix of the network.

7.2 Community detection

By a subpopulation in a structured population, we mean a group of individuals that share a common genetic history, and that are therefore more related to one another on average than they are to individuals outside the group. Individuals are then expected to share more genetic similarity edges within subpopulations than between subpopualtions. In network science, groups of nodes that are strongly connected within the group, forming dense subnetworks, are termed "communities" [14]. By detecting the community structure of genetic similarity networks, the genetic population structure can be revealed [9].

Many algorithms for community detection have been developed, including for weighted networks for which weights are assigned to the edges [16, 44]. Many of these methods are based on a quality function that evaluates community partitions, termed "modularity" [15]. This function, given a network and a partition of the network, returns a value between -1 and 1. Modularity is computed by analytically comparing the intra-community densities of a partition of the network with the expectation of intra-community densities over an implicit distribution of "random networks" that preserve the node weighted-degrees ([15]; Eq. 1 in [9]). Therefore, high modularity scores reflect a community partition in which within-community connections are more frequent and more highly-weighted than would be expected in a random network with the same degree distribution. Many community-detection algorithms aim to efficiently identify a community partition that maximizes the modularity score. Here, we use a particularly time-efficient weighted community-detection method, the Louvain method [45], which attempts to maximize modularity through a greedy search using a deterministic iterative process.

7.3 Edge-pruning

In structured populations, genetic variants are more likely to be shared within subpopulations than between subpopulations. Different genetic processes might have occurred over different spatial scales, either at different points in time or simultaneously, and therefore, the genetic signature is a complex aggregate of structure at several scales. Fine-scale population structure is characterized by groups of individuals who are strongly related, and that therefore are expected to be represented in subnetworks that contain only the strongest genetic similarity edges. Coarse-scale population structure, on the other-hand, is characterized by both strong and weak genetic similarities.

In order to detect population structure at multiple hierarchical levels, we perform an iterative edge-pruning process, where we sequentially remove edges below a genetic-similarity threshold. Here, we start at threshold 0 (i.e. all edges are included), and calculate the community structure. If only a single community is detected (no community structure), then we increment the threshold by 0.0001, remove the edges below the threshold, and again apply community detection. This threshold increment was selected for the analysis of the *A. thaliana* data such that the final output, the PST, does not substantially change with each increment; for other datasets, we recommend inspecting the typical differences between the sorted weights of the edges, selecting a threshold larger than the typical gaps between consecutive edge weights, and testing several threshold-increments to identify stable outputs (over several orders of magnitude). Once community structure is detected, the network is split, and the process is repeated for each community independently. This process of sudividing communities forms a PST, where clusters closer to the root represent coarse-scale structure and clusters at the leaves represent fine-scale structure.

The PST generated in this way is entirely data-driven, with the number of fine-scale and coarse-scale clusters determined only from the genetic data. Note that the population structure could include small clusters containing relatively few individuals. In some cases, even clusters containing only a few individuals might be informative, since they might represent sparsely sampled population fragments or familial groups. In other cases, such small groups may not be of interest. We therefore impose a stopping condition on the process by specifying the minimum number of individuals in the cluster, where clusters below that size are dropped. Because whole-genome data can potentially generate a fine-scale level of structure, we keep these minimum sizes small, and drop clusters smaller than five individuals for the *A. thaliana* analysis.

7.4 Generating geographical PST maps

We use a coloring scheme to visualize population structure. This scheme represents fine-scale structure while accounting for the hierarchical topology of population structure, and it allows identification of relevant differentiation patterns (Fig. 2 and Fig. 3). For each coloring, we select a root cluster from the PST. This cluster defines the scale and focus of the coloring; the coloring is intended to visualize the branch leading from this root cluster to the leaves. We start with an interval $[0, 1]$ representing a color gradient (here we use the "Rainbow" color gradient as implemented in Mathematica [46], with each number in the interval representing a color on the gradient; see Fig. S3 for an explicit representation of this color scheme function), and assign it to the root cluster. We then evenly split the interval among all immediate daughter clusters of the selected root cluster. We repeat this process, where in each step, a sub-interval of the parental interval is inherited by daughter clusters. In other words, if a cluster C is assigned the interval $[a, b]$, and it has n daughter clusters C_1,...,C_n, the daughter cluster C_i is assigned the interval $[a + \frac{(b-a)i-1}{n}, a + \frac{(b-a)i}{n}]$. This method assigns smaller intervals to clusters as we advance toward the leaves.

The process is terminated when leaves have been reached on all sub-branches. We then assign a single color to each cluster, by taking the color associated with the midpoint of the interval assigned to that cluster. In other words, if cluster C is assigned the interval $[d, e]$, then C is assigned the color corresponding to $\frac{d+e}{2}$. In this way, fine-scale clusters that are far from the root cluster and that share a common parent cluster one level above will tend to be assigned "closer" colors than will clusters that share a parent cluster nearer to the root cluster.

Each individual can be assigned to multiple clusters in different hierarchical levels, and we assign a color to an individual based on the finest-scale cluster to which it is assigned (i.e. the cluster closest to a leaf). This color assignment of individuals is then used for plotting geographic maps, where for each individual, a point is plotted at its sampling location, colored in its assigned color.

In this way, the many fine-scale clusters that are detected in large genomic datasets can be presented simultaneously, while allowing the broader genetic patterns to be clear on each map and at each scale by the coloring scheme. For example, although in Figures 2 and 3 many colors are used, some of which are visually indistinguishable, the broad patterns of genetic differentiation can be easily comprehended. When structure at a finer-scale is of interest, the cluster-coloring scheme can be repeated while selecting different root clusters, thus focusing on different sub-branches of the PST. This method makes it possible to rescale the coloring scheme to the appropriate geographic scale, and finer-resolution genetic differentiation patterns become apparent (e.g. Fig. 3). It is important to note that the PST holds information on all hierarchical levels; the coloring strategy we suggest only aids with visualizing the relationship of

genetic differentiation patterns to geography.

All maps were plotted using Mathematica [46] GeoGraphics, which imports maps from Wolfram alpha [47].

7.5 Pre-processing of the *A. thaliana* dataset

We used data from the 1001 Genomes project, which includes whole-genome SNP data of 1,135 *A. thaliana* diploid individuals (accessions) from Eurasia and North America [5], and data from Durvasula et al. [26] on 79 individuals from Africa. For the combined dataset of 1,214 individuals, we considered only biallelic SNPs, and filtering included removal of all SNPs with more than 10% missing data and all SNPs with minor allele frequencies <0.01. After filtering, the combined dataset included 1,214 individuals and 2,367,560 SNPs.

7.6 Comparing PSTs using NMI

To compare two PSTs derived from the same set of individuals I, we compute the normalized mutual information (NMI) between representations of the two PSTs as sets of partitions of I [17, 48]. NMI quantifies the amount of information shared by two random variables, or the amount of uncertainty on one variable that is reduced by observing the other variable, as measured in bits. In order to apply this measure, we represent a PST, X, with k clusters, as a partition X_1,...,X_k of the set of individuals I ($X_1 \cup X_2 \cup ... \cup X_k = I$). The sets X_1,...,X_k need not necessarily be mutually disjoint, and indeed in PSTs, clusters can be nested within other clusters (e.g., a daughter cluster is nested in the parent cluster). The NMI measure computes, for two PSTs X and Y represented as partitions, a score between 0 and 1, where higher scores represent higher similarities between the PSTs. We summarize the computation in Supplementary Information, and refer the readers to Lancichinetti *et al.* [17] and McDaid *et al.* [48] for the full details of this computation.

In some cases, it is of interest to compute NMI not for the entire PST, but rather for a subset of the clusters. For example, we computed NMI for leaf clusters of PSTs in order to compare the fine-scale structure described by the PSTs. Given two PSTs, X and Y, represented as partitions X_1,...,X_k and Y_1,...,Y_m, this computation is performed by considering only a subset of the sets in the partition, and computing NMI using only these subsets. However, because NMI requires both PSTs to be defined on the same set of individuals, a problem arises when the unions of the subsets of clusters are different. In other words, if we consider a subset of $k' \leq k$ of clusters from PST X and a subset of $m' \leq m$ clusters from PST Y, we might have $X_{a_1} \cup X_{a_2} \cup ... \cup X_{k'} \neq Y_{b_1} \cup Y_{b_2} \cup ... \cup Y_{m'}$. This problem can be addressed by removing or adding individuals to clusters in a standardized way, so that the unions

of the subsets of clusters become equal; our analysis of information gain by increased sequencing in *A. thaliana* provides an example of how such a case can be addressed.

7.7 Information gain by increased sequencing in *A. thaliana*

The NMI measure quantifies the information shared between two PSTs. If one of the PSTs considered was ascertained using strictly lower-density genetic data, then NMI can be used to evaluate the information gained by increasing the genomic density of the data, as we have done here for the *A. thaliana* dataset.

We calculated NMI for PSTs constructed from a subsample of the genome to simulate different fractions of the genome sequenced (0% to 100% in intervals of 2% of the number of SNPs in the dataset). First, the data were randomly partitioned to 10,000 approximately evenly-sized groups of SNPs, subsampled randomly from all SNPs without replacement, regardless of position in the genome. For each fraction of the genome, f, 100 random subsamples of the genome, each summing up to a fraction f of the genome, were taken without replacement (constructed using the SNP groups, for computational efficiency). For each such subsample, the PST was derived, and the NMI between that PST and the PST derived from the entire genome (Fig. 2A) was computed. The mean NMI and standard deviations for each f were then calculated.

In addition, we also considered the information gained when considering only fine-scale structure. This step used a similar process, only that the NMI was calculated by considering only the leaf clusters of the PSTs. Because the union of the leaf clusters does not necessarily cover the entire population, due to some individuals possibly not being assigned to any leaf, each individual not assigned to any leaf was added to all the leaves in the branch below the finest-scale cluster to which that individual was assigned. In other words, for PSTs X and Y, with leaf clusters $X_1,...,X_r$ and $Y_1,...,Y_s$, respectively, we add each individual d in $\bigcap_{i=1}^{r} X_i^c$ (where X_i^c is the complement of X_i) to all leaf clusters that belong to the branch emerging from the cluster $\bigcap_{x \in X | d \in x} x$; we do the same for all individuals in $\bigcap_{i=1}^{s} Y_i^c$ (see Supplementary Information for an illustrative example). We compute NMI for the sets $X_1,...,X_r$ and $Y_1,...,Y_s$ with the added individuals, which now cover the full set of individuals. This procedure generates a comparison between the finest-scale structures described by the PSTs.

7.8 Computational efficiency

A major issue with analyses of large genomic datasets is computational efficiency. The method we present requires two stages: (1) construction of the genetic similarity

network, and (2) construction of the PST. The first stage amounts to construction of a pairwise matrix of size n^2, where n is the number of individuals. Each element requires ℓ calculations, where ℓ is the number of loci. To apply Eq. 1, calculation of allele frequencies for all loci is required. The computation time for the first stage is therefore $O\big(ln^2 + ln\big) = O\big(ln^2\big)$ time. This computation time can be reduced substantially by computing the elements of the matrix for each of the loci in parallel: given c computation threads, the computation time can be reduced to $O(\frac{ln^2}{c})$ time.

The computation time of the second stage depends on the topology of the PST. With the community-detection Louvain algorithm, each community detection computation takes $O\big(i \log i\big)$ time, where i is the number of individuals in the network examined [45]. Therefore, for each level of the hierarchy, assuming we have k clusters in that level where cluster j is of size i_j, the computation time is $O\big(\sum_{j=1}^{k} i_j \log i_j\big)$ for that hierarchical level. Because $\sum_{j=1}^{k} i_j \leq n$ and $i_j \leq n$ (and therefore $\log i_j \leq \log n$) for any j, the run time of each level is bounded above by $O\big(n \log n\big)$. For a PST with d levels, the overall computational time of this step is, therefore, smaller than $O\big(dn \log n\big)$. The factor d is not known *a priori*, and could differ for different datasets; d is bounded by $d \leq n$, although in practice it would typically be much smaller.

In practice, for whole-genome data, the first step — constructing the matrix — takes substantially more time to compute than the second stage — iterative edge-pruning and community detection — because the second step does not depend on the number of loci in the data. For the *A. thaliana* dataset analyzed here, using a single computation thread on a university HPC cluster, the first step would require approximately 82 days. We used 1,000 computation threads on an HPC cluster, and the computation took 2 hours overall, emphasizing the importance of parallelizability in whole-genome population structure analyses. The second step required 49 minutes using a single computation thread.

References

1. Guillot, G., Leblois, R., Coulon, A. & Frantz, A. C. Statistical methods in spatial genetics. *Molecular Ecology* **18,** 4734–4756 (2009).
2. Allendorf, F. W., Hohenlohe, P. A. & Luikart, G. Genomics and the future of conservation genetics. *Nature Reviews Genetics* **11,** 697–709 (2010).
3. Novembre, J. & Peter, B. M. Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics and Development* **41,** 98–105 (2016).

4. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

5. Alonso-Blanco, C. *et al.* 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166,** 481–491 (2016).

6. Martin, H. C. *et al.* Insights into platypus population structure and history from whole-genome sequencing. *Molecular Biology and Evolution* **35,** 1238–1252 (2018).

7. Dyer, R. J. & Nason, J. D. Population graphs: the graph theoretic shape of genetic structure. *Molecular Ecology* **13,** 1713–1727 (2004).

8. Rozenfeld, A. F. *et al.* Network analysis identifies weak and strong links in a metapopulation system. *Proceedings of the National Academy of Sciences* **105,** 18824–18829 (2008).

9. Greenbaum, G., Templeton, A. R. & Bar-David, S. Inference and analysis of population structure using genetic data and network theory. *Genetics* **202,** 1299–1312 (2016).

10. Han, E. *et al.* Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications* **8,** 14238 (2017).

11. Greenbaum, G. & Fefferman, N. H. Application of network methods for understanding evolutionary dynamics in discrete habitats. *Molecular Ecology* **26,** 2850–2863 (2017).

12. Kuismin, M. O., Ahlinder, J. & Sillanpää, M. J. CONE: Community oriented network estimation is a versatile framework for inferring population structure in Large scale sequencing data. *G3: Genes—Genomes—Genetics* **7,** 3359–3377 (2017).

13. Newman, M. E. J. *Networks: An Introduction* (Oxford University Press, Oxford, UK, 2010).

14. Newman, M. E. J. in *Handbook of Graphs and Networks: From the Genome to the Internet* (eds Bornholdt, S. & Schuster, H. G.) 34–68 (Wiley-VCH, 2002).

15. Newman, M. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103,** 8577–8582 (2006).

16. Fortunato, S. Community detection in graphs. *Physics Reports* **486,** 75–174 (2010).

17. Lancichinetti, A., Fortunato, S. & Kertész, J. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics* **11,** 033015 (2009).

18. Sharbel, T. F., Haubold, B. & Mitchell-Olds, T. Genetic isolation by distance in *Arabidopsis thaliana*: Biogeography and postglacial colonization of Europe. *Molecular Ecology* **9,** 2109–2118 (2000).

19. Jørgensen, S. & Mauricio, R. Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured. *Molecular Ecology* **13,** 3403–3413 (2004).

20. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3,** 1289–1299 (2005).

21. Bakker, E. G. *et al.* Distribution of genetic variation within and among local populations of *Arabidopsis thaliana* over its species range. *Molecular Ecology* **15,** 1405–1418 (2006).

22. François, O., Blum, M. G. B., Jakobsson, M. & Rosenberg, N. A. Demographic history of European populations of Arabidopsis thaliana. *PLoS Genetics* **4,** e1000075 (2008).

23. Platt, A. *et al.* The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics* **6,** e1000843 (2010).

24. Provart, N. J. *et al.* 50 years of Arabidopsis research: Highlights and future directions. *New Phytologist* **209,** 921–944 (2016).

25. Lee, C. R. *et al.* On the post-glacial spread of human commensal Arabidopsis thaliana. *Nature Communications* **8,** 14458 (2017).

26. Durvasula, A. *et al.* African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* **114,** 5213–5218 (2017).

27. Hagmann, J. *et al.* Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genetics* **11** (2015).

28. Manel, S., Berthier, P. & Luikart, G. Detecting wildlife poaching: Identifying the origin of multilocus genotypes. *Conservation Biology* **16,** 650–659 (2002).

29. Rosenberg, N. A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1,** e70 (2005).

30. Morin, P. A., Martien, K. K. & Taylor, B. L. Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources* **9,** 66–73 (2009).

31. Haasl, R. J. & Payseur, B. A. Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity* **106,** 158–171 (2011).

32. Landguth, E. L. *et al.* Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Molecular Ecology Resources* **12,** 276–284 (2012).

33. Bryc, K., Bryc, W. & Silverstein, J. W. Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theoretical Population Biology* **89,** 34–43 (2013).

34. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).

35. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19,** 1655–1664 (2009).

36. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS genetics* **8,** e1002967 (2012).

37. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics* **48,** 94–100 (2015).

38. Li, Y., Huang, Y., Bergelson, J., Nordborg, M. & Borevitz, J. O. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences* **107,** 21199–21204 (2010).

39. Fournier-Level, A. *et al.* A map of local adaptation in *Arabidopsis thaliana*. *Science* **334,** 86–89 (2011).

40. Horton, M. W. *et al.* Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics* **44,** 212–6 (2012).

41. Busoms, S. *et al.* Salinity is an agent of divergent selection driving local adaptation of Arabidopsis to coastal habitats. *Plant Physiology* **168,** 915–929 (2015).

42. Long, Q. *et al.* Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics* **45,** 884–890 (2013).

43. Huber, C. D., Nordborg, M., Hermisson, J. & Hellmann, I. Keeping it local: Evidence for positive selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution* **31,** 3026–3039 (2014).

44. Yang, Z., Algesheimer, R. & Tessone, C. J. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports* **6,** 30750 (2016).

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

46. Wolfram Research, I. Mathematica, Version 11.3.

47. Wolfram Alpha. *Wolfram Alpha LLC* 2018.

48. McDaid, A. F., Greene, D. & Hurley, N. Normalized Mutual Information to evaluate overlapping community finding algorithms. *arXiv,* 1110.2515v2 (2013).