Supplementary Material: Data

New samples: We collected a total of 202 samples from nine populations that were genotyped on two different platforms (see Supplementary Table 1 below for details). In our sample collection process, we extracted DNA from blood samples of individuals that were at least 70 years old and self-reported that all four grandparents originated from the target population. We also took care to avoid relatives in our sampling. We expect that due to our sample selection process our data reflect the genetic structure of the target populations four generations before present.

Supplementary Table 1: New samples collected and genotyped in our work. Sample size indicates the number of individuals in our study.

Population	Region	Sample Size (N)	Genotyping Platform
Cappadocia	Anatolia	10	Illumina OMNI 2.5 (2,379,855 SNPs)
Crete	South Europe	90	Illumina OMNI1-QUAD (1,134,514 SNPs)
Dodecanese	South Europe	10	Illumina OMNI 2.5 (2,379,855 SNPs)
East Rumelia	South Europe	12	Illumina OMNI 2.5 (2,379,855 SNPs)
Macedonia	South Europe	16	Illumina OMNI 2.5 (2,379,855 SNPs)
Peloponnese	South Europe	19	Illumina OMNI 2.5 (2,379,855 SNPs)
Serbia	South Europe	20	Illumina OMNI 2.5 (2,379,855 SNPs)
Sicily	South Europe	20	Illumina OMNI 2.5 (2,379,855 SNPs)
South East Laconia	South Europe	5	Illumina OMNI 2.5 (2,379,855 SNPs)

Additional Datasets: We combined our data with four additional datasets in order to study population structure around the Mediterranean basin as well as Northern Europe. More specifically, we used data from ***(i)*** the Human Genome Diversity Panel (HGDP) [45], ***(ii)*** the HapMap Phase III Project [46], ***(iii)*** publicly available data on Northern African populations that were first released by Henn *et al.* 2012 [47], and ***(iv)*** data from the Kidd Lab at Yale University (allele frequencies for data from the Kidd Lab are available via the ALFRED database at http://alfred.med.yale.edu/). Supplementary Table 2 presents a detailed list of the 762 additional samples from 23 populations that were used in our main analyses.

Supplementary Table 2: An additional 762 samples from four other datasets were included in our main analyses. The genotyping platforms for each of the four additional datasets are described in the respective references. Sample size indicates the number of individuals in our study.

Population (country of origin)	**Region**	**Sample Size (N)**	**Reference**
Algeria	North Africa	19	Henn *et al 2012*
Egypt	North Africa	19	Henn *et al.2012*
Libya	North Africa	17	Henn *et al.2012*
North Morocco	North Africa	18	Henn *et al.2012*
Sahara Occidental	North Africa	18	Henn *et al.2012*
South Morocco	North Africa	16	Henn *et al.2012*
Tunisia	North Africa	18	Henn *et al.2012*
Bedouin	South West Asia	47	HGDP
Druze	South West Asia	46	HGDP
Mozabite	South West Asia	28	HGDP
Palestinian	South West Asia	51	HGDP
Yemen	South West Asia	37	Kidd Lab
Basque	South Europe	44	Henn *et al.2012,* HGDP
Italy	South Europe	13	HGDP
Sardinia	South Europe	28	HGDP
Tuscany	South Europe	96	HAPMAP, HGDP
Hungary	Central Europe	41	Kidd Lab
Chuvash	East Europe	41	Kidd Lab
Russia	East Europe	46	HGDP, Kidd Lab
France	West Europe	29	HGDP
Danes	North Europe	43	Kidd Lab
Finns	North Europe	25	Kidd Lab
Ireland	North Europe	22	Kidd Lab

Supplementary Figure 1: Geographic map showing the locations of the populations in Supplementary Tables 1 and 2 (see also Figure 1A in the main text).

Combined Datasets: The main dataset that was used in all our analyses emerged after combining the populations in Supplementary Tables 1 and 2. After extracting common SNPs, merging and aligning genotypes, and applying basic quality control filters (see Supplementary Methods for details), this dataset consisted of 964 samples from 32 populations, genotyped on 75,194 SNPs.

Supplementary Table 3: Approximate location of origin and geographic coordinates (longitude/latitude) for populations in our sample that surround the Mediterranean basin.

Population	**Capital**	**Latitude (North-South)**	**Longitude (East-West)**
Sahara Occidental	Laayoune	27.166695	-13.196338
South Morocco	Agadir	30.439202	-9.603809
North Morocco	Rabat	34.016242	-6.824268
Tunisia	Tunis	36.81808	10.149609
Mozabite	Ghardaia	32.488914	3.690967
Algeria	Algiers	36.738884	3.030469
Libya	Benghazi	32.129105	20.059277
Egypt	Cairo	30.040566	31.232373
Druze	Damascus	33.513919	36.308057
Palestine	Gaza	31.517679	34.452246
Cappadocia	Kayseri	38.73641	35.478864
East Rumelia	Plovdiv	42.140241	24.746992
Serbia	Belgrade	44.826169	20.463986
Macedonia	Thessaloniki	40.642875	22.944548
Peloponnese	Tripoli	37.511088	22.372785
SE Laconia	Neapoli Laconias	36.512844	23.060167
Dodecanese	Rhodes	36.435923	28.217683
Crete	Heraclion	35.340895	25.144416
Sicily	Catania	37.512722	15.083487
Italy	Rome	41.899722	12.481046
Tuscany	Florence	43.77506	11.247885
Sardinia	Olbia	40.927392	9.495078
Basque	Bilbao	43.265706	-2.934407

Supplementary Material: Methods and Results

Merging genotypes from different sources: In order to merge datasets from the five different sources described in Supplementary Tables 1, 2, and 3, we had to pay particular attention to strand information and properly align common SNPs. Since such information was not always available, we chose to omit SNPs with reference alleles CG and AT in order to avoid ambiguity.

Quality control: Despite the fact that the missing genotype rates in any of the five datasets that we analyzed in our work were quite low (invariably below 1%), we chose to perform an additional quality control check and remove any SNP that had a missing rate exceeding 20% in any of the populations under study. Our objective was to remove SNPs that might cause spurious artifacts in our analyses simply because they had many missing genotypes in one of the studied populations.

Computing Fst: We computed Fst between all available populations using the SmartPCA tool from the Eigenstrat software package, as well as our own MatLab script that implements the Fst formula of the International HapMap Project.

PCA: We used our own MatLab implementation of PCA, which mean-centers the data while ignoring missing entries. We first transformed the genotypic data to numeric values, without any loss of information, in order to apply linear algebraic analyses. Consider a dataset for a group of populations consisting of m subjects and assume that for each subject n biallelic SNPs have been assayed. Thus, we are given a table T, consisting of m rows and n columns. Each entry in the table is a genotype (pair of bases), ordered alphabetically. We transform this initial data table to an integer matrix A, which consists of m rows (one for each subject) and n columns (one for each SNP). Each entry of A will be set to +1, 0, -1; an entry could be empty in the case of missing genotypes. Let B_1 and B_2 be the bases that appear in the j-th SNP (in alphabetical order). If the genotypic information for the j-th SNP of the i-th individual is B_1B_1 the (i,j)-th entry of A is set to +1; else if it is B_1B_2 the (i,j)-th entry of A_X is set to 0; else if it is B_2B_2 the (i,j)-th entry of A is set to -1; see Paschou et al. 2007a [48], Paschou et al. 2007b [49] for more details. After forming the matrix A, we transform it to a mean-centered matrix M, by mean-centering each column (SNP) of A. It is worth noting that we ignore missing entries when forming M. Finally, in order to compute the principal components, it suffices to compute the Singular Value Decomposition of the matrix product MM^T, which is an m-by-m matrix.

Supplementary Figure 2 (five panels) shows PCA plots (in two or three dimensions) for subsets of the populations of Supplementary Tables 1 and 2. Panels (a) and (B) are PCA plots of all populations in Supplementary Tables 1 and 2, projected on the top two and three principal components. Panel (c) is the three-dimensional analog of main text Figure 1b. Panels (d) and (e) are PCA plots of populations around the Mediterranean basin, with Bedouin and Yemenites excluded. There is an obvious resemblance between the geography of the region and the two-dimensional PCA plots.

Different sample sizes and their effect on Principal Component Analysis: To test the effect of different sample sizes on PCA, we computed centroids of our input populations in Supplementary Tables 1 and 2 first using all available samples, then using only 20 samples per population (chosen uniformly at random), and finally using only 30 samples per population (chosen uniformly at random). We then correlated the centroids of the populations using all available samples with the centroids that were computed using 20 and 30 samples only. The

Pearson correlation coefficient for the first principal component between the centroids that were computed using all available samples and the centroids that were computed using 30 (respectively 20) samples per population exceeded 0.956 (respectively 0.955) with a standard deviation (over 100 repetitions) of less than 0.001 (resp. 0.001). The correlation for the second principal component was only slightly worse in the case of 20 samples per population: the Pearson correlation coefficient for the second principal component between the centroids that were computed using all available samples and the centroids that were computed using 30 (respectively 20) samples per population exceeded 0.93 (respectively 0.85) with a standard deviation (over 100 repetitions) of less than 0.006 (respectively 0.02). Given these statistics, we opted to keep all available samples in our populations in order to capture as much of the population variance as possible.

Correlation between PCA and geographic coordinates: In order to estimate the correlation between geographic coordinates and the top two eigenvectors emerging from PCA, we used the data in Supplementary Table 3. For each population in our sample, we approximated its location of origin either using information provided to us by the individuals that collected the respective sample, or by using a capital city that is relatively close to the population under study. The correlation between geographic coordinates and the eigenvectors was computed by converting both the geographic coordinate vector and the eigenvectors to z-scores, and then computing the Pearson correlation coefficient. We also ran a Mantel test in order to estimate the statistical significance of the result by permuting the geographic distance matrix between all pairs of populations (we note that we used the Haversine formula to accurately compute distance between populations using longitude and latitude information). After 10,000 permutations we did not get a single permutation whose Pearson correlation coefficient exceeded the observed ones, thus concluding that our observation is significant with a p-value at least 10^{-5}.

Identification of genetic barriers: We used the BARRIER v2.2 software [26] to explore the genetic barriers in our dataset. BARRIER implements Monmonier's maximum difference algorithm in order to identify such barriers. First, using the geographic coordinates of the 23 populations surrounding the Mediterranean sea (see Supplementary Figure 2d for the PCA plot of these populations), we used BARRIER to plot the studied populations on the two dimensional plane (see Supplementary Figure 3a). Supplementary Figure 3a also shows the Delaunay triangulation and the Voronoi tessellation for the 23 populations. Recall that the Voronoi tessellation (blue lines) can be thought of as defining polygons that include all points on the plane that are closer to the centroid of the polygon under investigation than to the centroid of any other polygon. Since the centroids of the polygons in BARRIER correspond to the studied populations, we can consider the respective polygons as the areas in the plane that correspond to the respective population. If two polygons (and their respective populations) share an edge, then one might conclude that, based on the geographic distances, the respective populations are connected and thus some gene flow between the two populations could be expected. Indeed, the green edges in Supplementary Figure 3a (corresponding to the Delaunay triangulation) indicate connections between various pairs on populations based on geography; clearly, many connections exist that cross the Mediterranean sea: note that populations 1, 10, 12, 13, 14, 18, 22 all correspond to North African populations, yet, based solely on geography, they have multiple connections with South European populations.

This picture changes drastically once genetic information from Principal Components Analysis is incorporated. In order to get statistical significance, we used a standard bootstrap procedure to

compute 100 distance matrices by resampling individuals from our populations. To be precise, we used a ten-fold cross-validation procedure in order to construct these 100 distance matrices, where each matrix contains all pairwise distances between the centroids of the 23 studied populations, as computed via Principal Components Analysis on the sampled individuals. The BARRIER software accepts the bootstrap distance matrices as input in order to run Monmonier's algorithm in order to compute genetic barriers between the studied populations. Supplementary Figure 3b clearly shows (observe the thick red lines) strong genetic barriers separating the North African populations from the South European populations, as expected by the presence of the Mediterranean sea.

Estimating population admixture: We used the ADMIXTURE v1.22 software for all our admixture analyses. The parameter K (number of ancestral populations) ranged between two and eight in all our analyses. Prior to running ADMIXTURE, we pruned the SNPs in order to remove SNPs in high LD. Towards that end, we used the PLINK software and pruned SNPs using a windowed approach and a value of r^2 equal to 0.8. We used DISTRUCT v.1.1 and CLUMPP v.1.1.2 to visualize the output of ADMIXTURE.

Supplementary Figure 4 shows an ADMIXTURE plot of all populations in Supplementary Tables 1 and 2, for all values of K between two and eight. Supplementary Figure 5 shows an ADMIXTURE plot of all populations in Supplementary Tables 1 and 2, excluding the Bedouins and the Yemenites, for all values of K between two and eight. The data in Supplementary Figure 5, namely the percentages of origin of each individual with respect to the K (unknown) ancestral populations, were used in our ADMIXTURE based network formation algorithms (see Supplementary Figure 8).

In order to account for potential variance in multiple runs of ADMIXTURE, we performed 20 runs of ADMIXTURE for each value of K between two and eight. We focus our discussion on the dataset used in main text Figure 3; the conclusions for the datasets used in Supplementary Figures 4 and 5 are identical. Each of the 20 runs (for a fixed value of K) was performed with a different random seed using the "–s time" parameter of ADMIXTURE in order to initialize the random seed generator using the current time of the machine; this is often considered to be a good approximation to true randomness. We observed that for all values of K the results were very highly correlated. More specifically, recall that ADMIXTURE returns, for each sample in the dataset, K coefficients indicating percentage of ancestry in each of the K (assumed) ancestral populations. First of all, for a particular value of K, the average Pearson correlation coefficient among the reported ADMIXTURE coordinates and the coordinates returned in the 20 replicates ranged from 0.94 for K=2 to 0.86 for K=8. This already indicates that the correlation between the various replicates of ADMIXTURE is very high (a similar observation was made by [Henn et al. (2011) Proc Natl Acad Sci USA 108(13):5154-5162], who reported that multiple runs of ADMIXTURE only resulted in minor changes and simply chose to report one run; in a subsequent paper, [47] only performed a single run of ADMIXTURE). The situation is actually even more favorable, once we look closer at the data: more specifically, for each of the 20 runs (say run j, for some j between one and 20), we computed the Pearson correlation coefficient between the reported coefficients for each of the K ancestral populations (as shown in main text Figure 3) and their best fit in the j-th run (we implemented a matching algorithm to perform the alignment between the K ancestral populations in the reported run and the j-th replicate). We noticed that, for all values of K between two and six, over 70% of the replicates were essentially identical to the reported one (average Pearson correlation coefficient exceeded 0.98), while for K equal to seven or eight approximately 50% of the replicates were essentially identical to the reported one (average Pearson correlation coefficient approximately 0.95). Checking the

reported values of log-likelihood and delta (as reported by ADMIXTURE), we observed that for the cluster of identical repetitions convergence was better than for the other repetitions, thus indicating a better approximation to the true optimal value. Therefore, we chose to report one of the essentially identical runs. A similar analysis was performed in [Behar et al. (2010) Nature (8) 466 (7303):238-42], who used a similar method to account for the slight variation in the ADMIXTURE output (and then chose to report one of the identical ADMIXTURE outputs).

f3 statistics: In order to test for bi-directional admixture and population splits along the populations that stand at the gateway to Europe, we ran three population tests as described by Patterson et al [27] and implemented in TreeMix [30]. Seven populations bridging Anatolia to Southern Europe were included in the analysis (Cappadocia, Dodecanese, Crete, SE Laconia, Peloponnese, Macedonia, East Rumelia), and f3 statistics were calculated for all possible triplets of populations. After removal of SNPs in LD (at $r^2>0.5$) from our initial dataset of over 650,000 SNPs, we were left with approximately 278,000 SNPs. When studying a triplet of populations (e.g., C; A,B) f3 statistics and the corresponding Z-score of significance can only be negative if population C has ancestry from populations related to both A and B. It is only in this case that paths exist between C and A as well as C and B that also take opposite drift directions. The observation of a significantly negative value of f3(C; A, B) is thus evidence of complex phylogeny in C. All triplets of populations are shown in our online supplementary material (http://www.cs.rpi.edu//~drinep/Maritime_Route/f3.xls). We only get marginally negative results for the Peloponnese population in Greece, indicating that this population might have resulted from some admixture from both North and South. All other tested triplets tested positive, indicating that there are simple paths connecting populations to each other along the course of migration from Anatolia into Europe. This is in concordance with our network analysis.

Network analysis: To better understand the connection between the populations included in our study we performed a network analysis on the results of PCA and ADMIXTURE.

We start by noting that both PCA and ADMIXTURE are essentially reducing the dimensionality of the original dataset by expressing each sample as either *(a)* a linear combination of the top few eigenvectors (in the case of PCA) or *(b)* as percentages of ancestry from a small number of – typically unknown – ancestral populations (in the case of ADMIXTURE). To be more precise, each individual sample is originally described with respect to 75,194 SNPs; mathematically, this is equivalent to saying that the sample lies in a 75,194-dimensional subspace. After applying PCA or ADMIXTURE to the dataset, all samples are described with respect to K coefficients; K ranges between one and eight in all our analyses. Mathematically, this is equivalent to saying that the output of PCA or ADMIXTURE lies in a K-dimensional subspace, with $K <<< 75{,}194$. It is worth highlighting that PCA and ADMIXTURE are unrelated dimensionality reduction techniques that make different assumptions regarding the underlying structure of the dataset. PCA is a very general *algorithmic* technique that recovers the latent structure of high-dimensional data that lie in an approximately linear manifold, by leveraging the existence of a small number of pairwise orthogonal principal components that capture the linear structure of the dataset. ADMIXTURE, on the other hand, is a much more specialized *model-based* statistical analysis technique whose underlying model, to the best of our understanding, is only applicable to population genetics datasets. Its end goal is a (maximum likelihood based) estimation of individual ancestries from large-scale SNP datasets. An important difference regarding the output of the two techniques is that the output of PCA lies in a low-dimensional *normed* subspace as described above, where the metric of distance between two samples is the Euclidean distance, since PCA minimizes an objective function that is based on the Euclidean or ℓ_2 norm. (Recall that the Euclidean or ℓ_2 distance between a pair of points in a K-dimensional space is the square root of the sum of the squares of the coordinate-wise distances between the

two points.) However, while the output of ADMIXTURE also lies in a low-dimensional subspace, the relevant norm is not necessarily the Euclidean norm. As a matter of fact, the output of ADMIXTURE is essentially a probability distribution for each sample, indicating the percentage of origin of the sample with respect to the ancestral populations. Indeed, the coefficients returned by ADMIXTURE for each sample are positive and sum up to one, a fact that does not hold for PCA. Thus, a more appropriate choice of norm for ADMIXTURE's output is the ℓ_1-norm, which better captures the distance between two probability distributions. (Recall that the ℓ_1 distance between a pair of points in a K-dimensional space is sum of the absolute values of the coordinate-wise distances between the two points.)

In order to form a network of related populations using the output of PCA or ADMIXTURE, we first identify the top ten nearest neighbors of each individual in the PCA or ADMIXTURE dimensionally reduced subspace, with the additional constraint that these nearest neighbors must not lie in the population of origin of the target individual. To be more precise, consider an individual in population X; this individual is represented with respect to K coefficients, which are the output of PCA or ADMIXTURE. We start by computing the ℓ_2 (Euclidean) distance in the case of PCA (or ℓ_1 distance in the case of ADMIXTURE) between the target individual and every other individual in our dataset. Then, we identify the top ten nearest neighbors to the individual under study, with the constraint that these nearest neighbors do not belong to population X. (We note that the distances are only used in order to select the top ten nearest neighbors from the target individual; the numeric values of the distances of those neighbors and the target individual are ignored in the remainder of our network analysis.) This procedure is repeated for all individuals in the dataset.

Assuming that we have *n* samples from *p* populations in our dataset, this procedure eventually returns an *n-by-p* table with rows corresponding to samples and columns corresponding to the populations under study. The (i,j)-th entry in the table denotes how many neighbors individual *i* has in population *j*. Clearly, by the construction of our nearest-neighbor identification algorithm, the (i,j)-th entry of the matrix is equal to zero if the j-th column of the table corresponds to the population of origin of the i-th sample.

Given the above *n-by-p* table A, we are now ready to create a network of populations with edges connecting related populations. In order to describe our procedure to infer whether populations X and Y are connected, as well as the weight of the respective connection, we first need to introduce some notation. Let t be the cardinality of population X and let s be the cardinality of population Y. Let $i_1, i_2, \ldots, i_t$ be the row indices corresponding to samples in X. Let j_Y be the column of A that corresponds to population Y and compute

$$dist(Y, X) = \sqrt{\frac{1}{s}\sum_{a=1}^{t} A_{i_a j_Y}}$$

Similarly, if we use $i_1, i_2, \ldots, i_s$ to denote the row indices of the table A that correspond to samples in Y (recall that the cardinality of Y is s) and j_X to denote the column of A that corresponds to population X, we can compute

$$dist(Y, X) = \sqrt{\frac{1}{t}\sum_{a=1}^{s} A_{i_a j_X}}$$

It should be obvious that, up to scaling, our distance metric simply counts the number of nearest neighbors that individuals of population X have in population Y. This is a simple and intuitive

idea; a minor fine tuning is necessary in order to mitigate the effect of different population sizes in our sample.

A number of comments will help the reader to better understand the above two metrics of distance between populations X and Y. First of all, our distance metric is not symmetric: mathematically, dist(X,Y) is not necessarily equal to dist(Y,X). Second, in order to mitigate the effect of different population sizes, we normalize the aforementioned count by dividing it by the square root of the population size. The goal of this normalization is to assign larger weights when nearest neighbors in smaller populations are identified. For example, identifying a nearest neighbor in a population Y that has only 10 samples should be weighted more than identifying a nearest neighbor in a population Y that has 100 samples. The choice of normalizing by the square root (more generally, a power less than one) of the sample size is a common tradeoff in order to mitigate the effects of different population sizes, while not overcompensating for such effects. Third, using the dimensionally reduced space (for various values of K) derived by PCA or ADMIXTURE in order to compute the nearest neighbors of each individual (instead of using all available genotypes and the allelic distance) has two advantages: first and foremost, it *denoises* the data by keeping only the most prominent and meaningful axes of variance and thus avoiding the statistical artifacts of the curse-of-dimensionality; second, it speeds up computations since the data now lie in a low-dimensional space.

We are now ready to describe our network formation algorithm. For each pair of populations X and Y, we compute both distances: dist(X,Y) and dist(Y,X). We create an edge between X and Y (thus claiming that the two populations are neighbors of each other) if

$$\min\{\text{dist}(X,Y),\text{dist}(Y,X)\} > 0,$$

and we assign as a weight of the respective edge the value min{dist(X,Y),dist(Y,X)}. It is worth noting that our choice is quite conservative, since we will assume that populations X and Y are related if and only if both X and Y have nearest neighbors in each other. This notion of edge weights handles situations where a population Y is far away from all other populations and thus dist(X,Y)=0 for all populations X, but, by construction, there will exist some populations X so that dist(Y,X)>0 (since each sample in Y will have some nearest neighbors in populations outside Y, even if Y is isolated).

Finally, once a network whose nodes correspond to populations and whose edges correspond to connections between populations, as described above, is formed, we visualize it using the Cytoscape software package.

We briefly discuss the number of nearest neighbors to be retained using our approach. It should be clear that our network formation algorithm has one free parameter, namely the number of nearest neighbors to retain in the first step of the process. We chose to set that value to ten, which is equal to 1/3 of the average number of samples per population in our dataset. (Supplementary Tables 1 and 2 show the number of samples per population; the average number of samples/population is equal to 30.1) We also experimented with the number of nearest neighbors set to eight, nine, eleven, and twelve, without observing noticeable differences in the resulting networks.

Our network analysis using PCA is based on Supplementary Figures 2d and 2e, as well as their higher dimensional analogs, which of course cannot be plotted. Note that we included all

populations in Supplementary Tables 1 and 2, except for the Bedouins and the Yemenites, which form a separate cline towards Central-South Asia. Similarly, our network analyses using ADMIXTURE were based on Supplementary Figure 5, which again excludes the Bedouins and the Yemenites.

First, in the case of PCA, we had to determine how many principal components are significant. Towards that end, we examined the distribution of singular values in Supplementary Figure 6 (both panels). Panel (a), which corresponds to the plots in Supplementary Figures 2d and 2e, shows that the singular values drop fast: the top three are clearly significant; then the fourth and the fifth one, as well as the sixth and the seventh one, form small clusters. A simple metric of significance (see the legend of Supplementary Figure 6) indicates that the top seven singular values are indeed significant. Adding the Bedouins and the Yemenites to our dataset results in a slight increase in the number of significant singular values, which is now equal to ten.

Supplementary Figures 7 and 8 show the various networks that were formed using all values of K and either PCA or ADMIXTURE as the basis for computing nearest neighbors. In the case of PCA, we examined values between three and seven (the number of significant principal components). In the case of ADMIXTURE, we examined values of K between three and eight. Interestingly, all networks are highly similar, showing a distinct path from North Africa to Near East and Southern Europe via Cappadocia, Dodecanese, and Crete. The resulting networks, as visualized by Cytoscape, are also available online as .cys files and include numerous network statistics, as computed by Cytoscape (see Supplementary Online Material).

Network analysis via Fst: We also performed a network analysis using the Fst metric as follows: we formed an edge between two populations if the respective Fst was below 0.08 (we chose this threshold value since it resulted in keeping the top 25% most significant Fst pairs while maintaining a connected network – larger threshold values resulted in similar yet denser networks, while smaller values disconnected the network). The resulting network is shown in Supplementary Figure 9 and is similar in spirit to the networks formed via our metric that is based on PCA and ADMIXTURE. We believe that PCA and ADMIXTURE based networks capture in a much finer sense the connections between the analyzed populations; Fst is a much coarser statistic.

Simulating a stepping-stone model: To further test the stepping-stone hypothesis regarding the migration of populations around the Mediterranean basin, we used IBDSim to generate simulated data. IBDSim [29] is a package for the simulation of genotypic data under isolation by distance. It is based on a backward ‘generation by generation’ coalescent algorithm allowing the consideration of various isolation-by-distance models. We used IBDSim in order to generate 10,000 independent SNPs for ten populations that were placed on a one-dimensional lattice on coordinates that are representations of a subset of ten populations from our data. We set the effective population size to 30, the mutation rate to 0.5e-6, and the migration rate to 0.5. (See Supplementary Table 4, where ten consecutive -- from a geographic perspective -- populations around the Mediterranean basin were chosen; the table also indicates their geographic coordinates, the distance from one population to the next one, the normalized distance from one population to the next one, and the respective lattice points where the populations were placed.

More specifically, we treated the minimum distance between any two consecutive populations in our list as a single unit; this was the distance between Crete and the Dodecanese. Then, we divided each distance by the minimum distance, thus getting the sixth column of the table. Finally, we placed populations in the x-axis of a one-dimensional lattice by rounding the distances in the sixth column to the nearest integer in order to approximately respect the distances between two consecutive populations.) We used default values for the migration rate and the mutation rate and generated data using the stepping stone model, as implemented by IBDSim. Recall that in the stepping stone model, an individual can only move to an adjacent population. The resulting PCA plots and the corresponding network are shown in Supplementary Figures 10 and 11 and they are clearly reminiscent of patterns presented in Supplementary Figures 2 and 7 (PCA and network analysis of real data). Thus our simulations provide further support for our conclusions regarding the stepping stone hypothesis of population migrations around the Mediterranean with Crete acting as a hub that connects Anatolia to Southeastern Europe.

Inferring phylogenetic trees: We used PHYLIP 3.695 to infer phylogenetic trees using the output of SmartPCA (Fst distances between all pairs of available populations). Phylogenetic trees were constructed using the neighbor-joining method (NEIGHBOR tool in PHYLIP). DRAWGRAM and DRAWTREE were used in order to plot the resulting phylogenetic trees.

In order to further verify our findings, we constructed a phylogenetic tree based on the Fst distance matrix. The resulting tree (see Supplementary Figure 12 that includes the Sub-Saharan African San population as an outgroup – this is a South African population that has been made available by the HGDP consortium) clearly support the findings of PCA, ADMIXTURE, as well as our network analyses.

TreeMix: To further verify the migration pathways and signals of admixture among the studied populations from Northern Africa, Middle East, Anatolia, and Europe, we additionally used TreeMix [30] to find a population graph that best describes the relationship between populations in the dataset by testing for gene flow between them. Genomewide allele frequency data is used to first find the maximum-likelihood tree of populations and then infer possible additional migration events by identifying populations that poorly fit this tree. The maximum likelihood tree of all populations included in this analysis is shown in Supplementary Figure 13a, with the respective residuals shown in Supplementary Figure 13b. (It is worth noting that this tree explains 96.3% of the variation of the data, as computed by TreeMix.) In concordance with all other analysis that we present, gene flow from Anatolia to Southern Europe appears to have occurred through the islands of the Dodecanese and Crete. Residuals were subsequently analyzed in order to identify pairs of populations that are more related to each other than is captured by this graph and corresponding migration events were inferred. We then sequentially added 10 migration events to reach a tree that captures approximately 98.5% of the data. Migration events are shown in Supplementary Table 5. TreeMix analysis supports some additional gene flow among populations from Northern Africa (e.g., from South Morocco to Egypt and Algeria), as well as a connection between South Morocco and Middle East. Furthermore, an indication of the complex genetic history of Sardinia is uncovered with connections to Northern Africa, Italy, and the Basques.

Neighbor-Net analysis: As a final verification of the migration pathways and signals of admixture among the studied populations, we performed a phylogenetic network analysis using the Neighbor-Net method [31] implemented in SplitsTree v4.13. Neighbor-Net is a distance based method for the construction of phylogenetic networks that is based on the neighbor-joining algorithm. Recall that a fundamental property of a tree is that between every pair of populations there is a unique path connecting them, while a network or *graph* allows for multiple paths. The Neighbor-Net algorithm relaxes the neighbor-joining algorithm to construct a phylogenetic graph that aims to improve the fit to the data by allowing (if necessary) additional paths between pairs of populations. Comparing to our own network formation algorithms, it is worth noting that our approach is not based on neighbor-joining and does not attempt to fit any particular structure to the data. It simply picks the most informative connections in order to produce a graph of gene flow among populations.

In order to run Neighbor-Net, we first computed the Fst matrix between all populations using EIGENSOFT 5.0.1 on our dataset and used this matrix as the input for the Neighbor-Net algorithm. The Neighbor-Net method was used to produce an unrooted network, which was subsequently visualized using the EqualAngle method (Supplementary Figure 14). In concordance with all of our previous analysis, it is again clear that the Dodecanese and Crete play a pivotal role as a hub in gene flow connecting Anatolia to Southern Europe, and the rest of Europe.

Supplementary Material: Figures

Supplementary Figure 2: PCA plots of (subsets of) the populations in Supplementary Tables 1 and 2. (a) PCA plot of all populations in Supplementary Tables 1 and 2; projection on top two Principal Components. The cline from Northern Africa to Northern Europe via Near East, Cappadocia, Dodecanese, and Crete is apparent. (b) PCA plot of all populations in Supplementary Tables 1 and 2; projection on top three Principal Components. (c) PCA plot of populations around the Mediterranean basin; projection on top three principal components. This plot is the three-dimensional version of main text Figure 1b. Bedouins and Yemenites are separated from the Mediterranean basin populations. (d) PCA plot of populations around the Mediterranean basin; projection on top two principal components. (Bedouin and Yemenites haven been excluded from this plot.) (e) PCA plot of populations around the Mediterranean basin; projection on top three principal components. (Bedouin and Yemenites have been excluded from this plot.) The Tunisian population is separated from the other populations in the third principal component.

Figure S2a

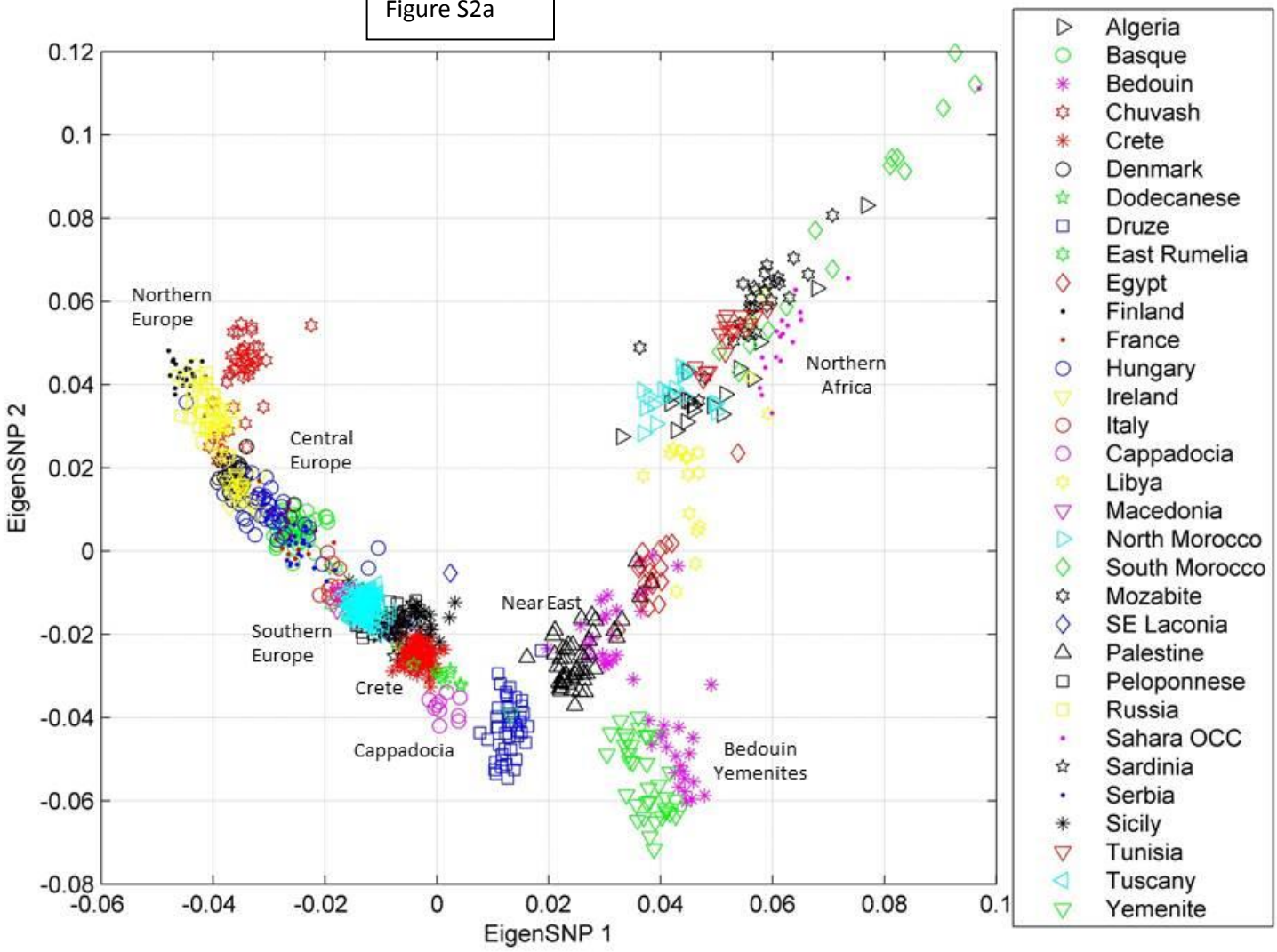

Northern Europe
Central Europe
Southern Europe
Crete
Cappadocia
Near East
Northern Africa
Bedouin Yemenites
EigenSNP 1
EigenSNP 2
Algeria
Basque
Bedouin
Chuvash
Crete
Denmark
Dodecanese
Druze
East Rumelia
Egypt
Finland
France
Hungary
Ireland
Italy
Cappadocia
Libya
Macedonia
North Morocco
South Morocco
Mozabite
SE Laconia
Palestine
Peloponnese
Russia
Sahara OCC
Sardinia
Serbia
Sicily
Tunisia
Tuscany
Yemenite

Figure S2b

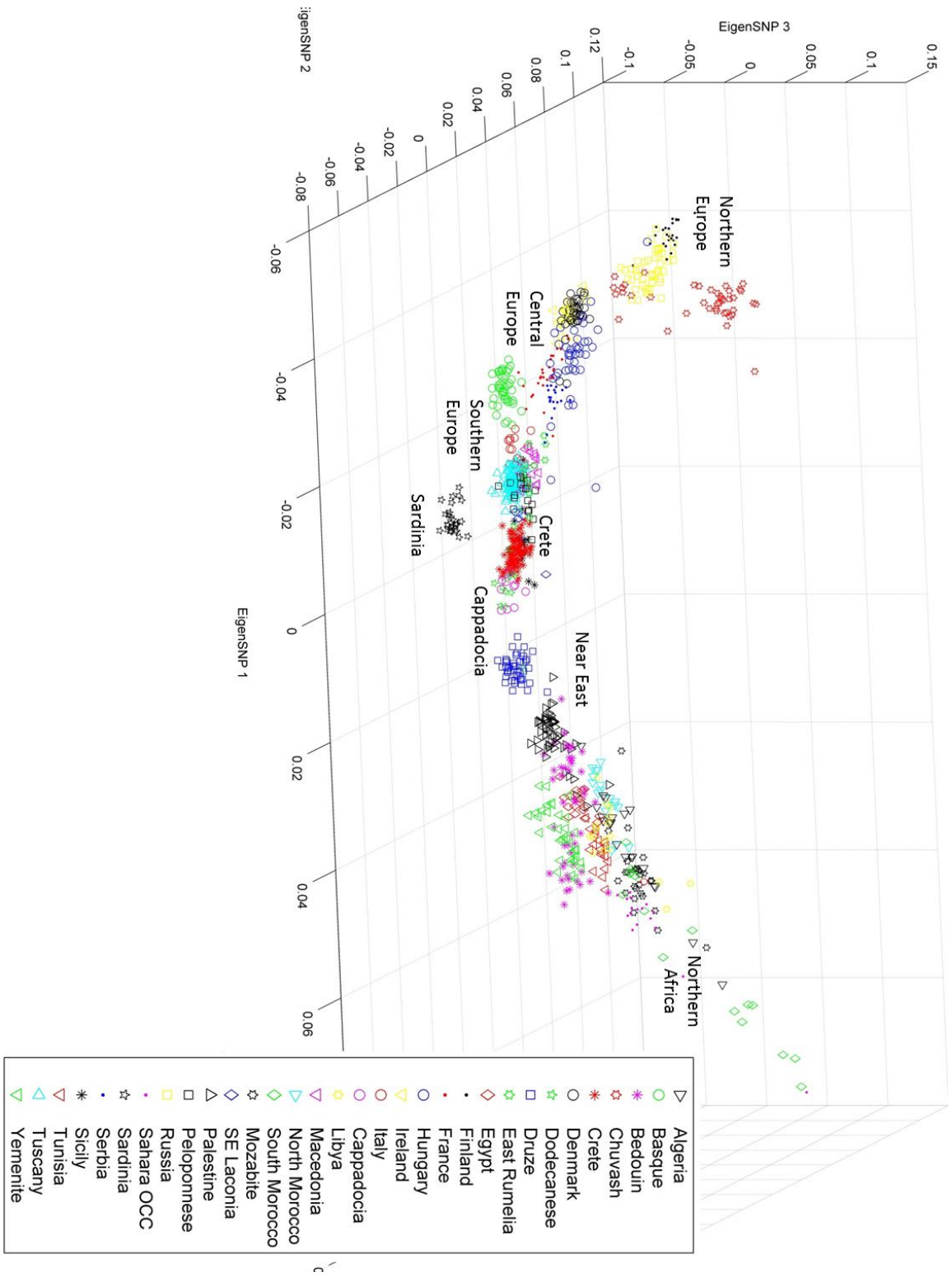

EigenSNP 3
EigenSNP 2
EigenSNP 1
Northern Europe
Central Europe
Southern Europe
Sardinia
Crete
Cappadocia
Near East
Northern Africa
Algeria
Basque
Bedouin
Chuvash
Crete
Denmark
Dodecanese
Druze
East Rumelia
Egypt
Finland
France
Hungary
Ireland
Italy
Cappadocia
Libya
Macedonia
North Morocco
South Morocco
Mozabite
SE Laconia
Palestine
Peloponnese
Russia
Sahara OCC
Sardinia
Serbia
Sicily
Tunisia
Tuscany
Yemenite

Figure S2c

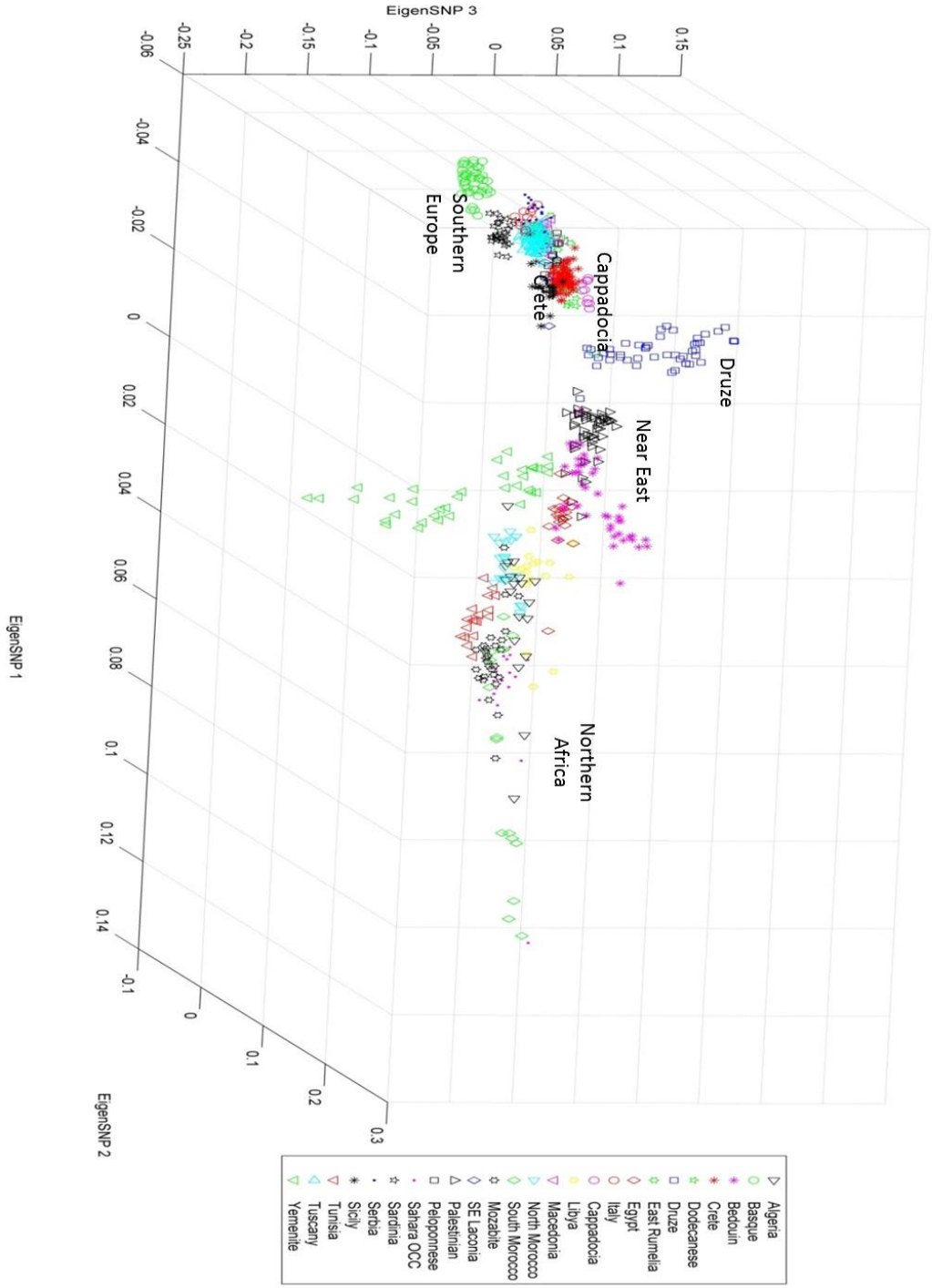

EigenSNP 3
EigenSNP 1
EigenSNP 2
Southern Europe
Cappadocia
Crete
Druze
Near East
Northern Africa
Algeria
Basque
Bedouin
Crete
Dodecanese
Druze
East Rumelia
Egypt
Italy
Cappadocia
Libya
Macedonia
North Morocco
South Morocco
Mozabite
SE Laconia
Palestinian
Peloponnese
Sahara OCC
Sardinia
Serbia
Sicily
Tunisia
Tuscany
Yemenite

Figure S2d

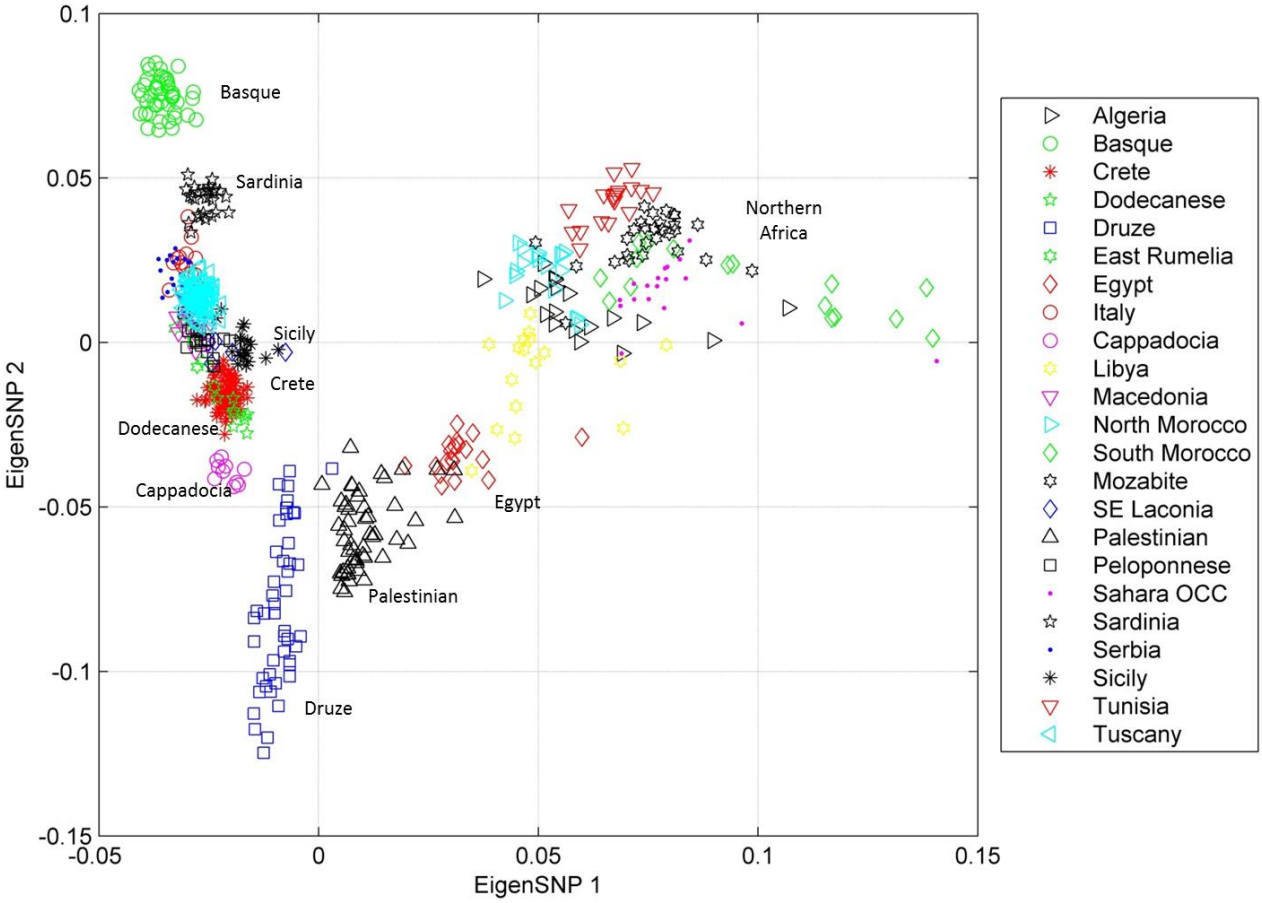
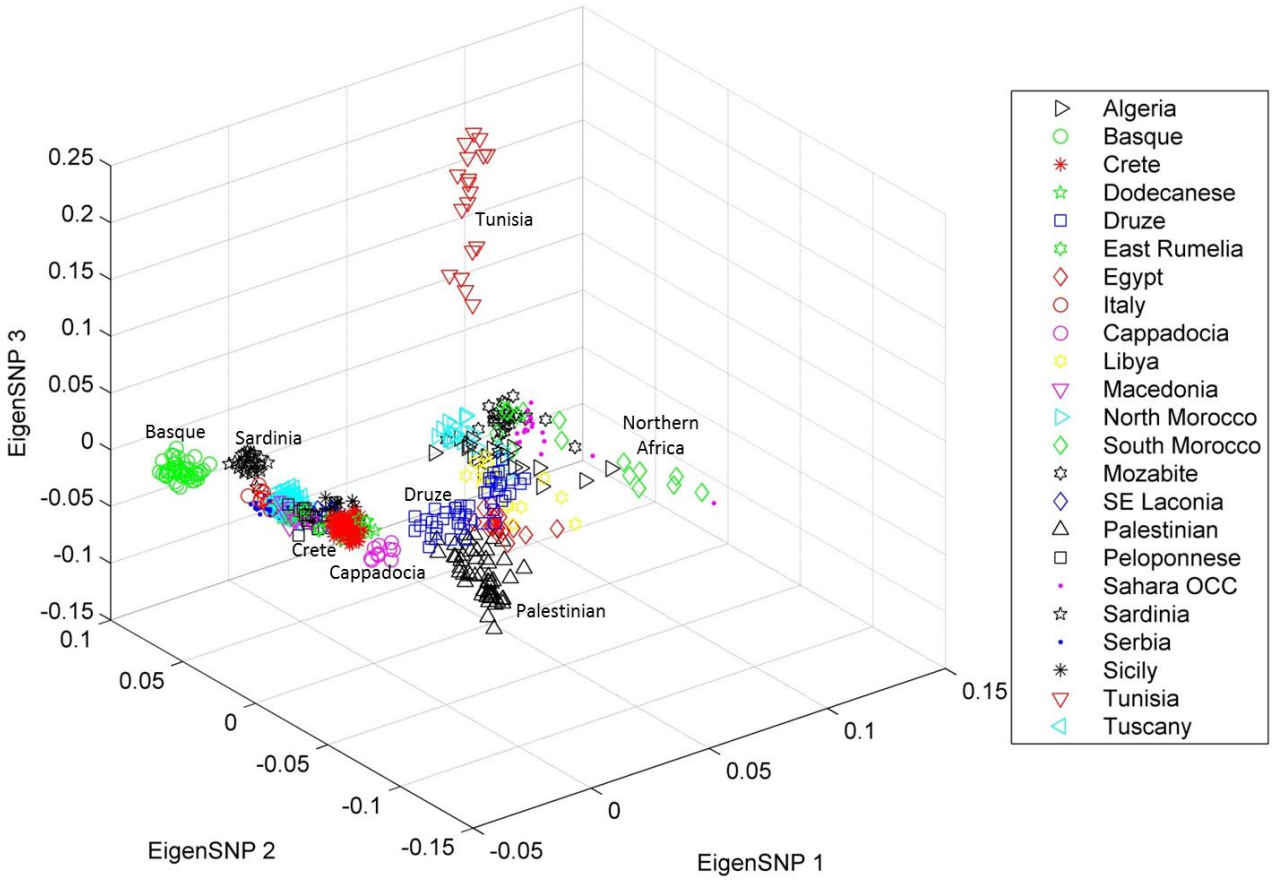

0.1
0.05
0
-0.05
-0.1
-0.15
-0.05
0
0.05
0.1
0.15
EigenSNP 1
EigenSNP 2
Basque
Sardinia
Sicily
Crete
Dodecanese
Cappadocia
Egypt
Palestinian
Druze
Northern
Africa
Algeria
Basque
Crete
Dodecanese
Druze
East Rumelia
Egypt
Italy
Cappadocia
Libya
Macedonia
North Morocco
South Morocco
Mozabite
SE Laconia
Palestinian
Peloponnese
Sahara OCC
Sardinia
Serbia
Sicily
Tunisia
Tuscany

Figure S2e


Tunisia
Basque
Sardinia
Druze
Crete
Cappadocia
Northern
Africa
Palestinian
EigenSNP 3
EigenSNP 2
EigenSNP 1
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
Algeria
Basque
Crete
Dodecanese
Druze
East Rumelia
Egypt
Italy
Cappadocia
Libya
Macedonia
North Morocco
South Morocco
Mozabite
SE Laconia
Palestinian
Peloponnese
Sahara OCC
Sardinia
Serbia
Sicily
Tunisia
Tuscany

Supplementary Figure 3: Results from running the BARRIER v2.2 software. (a) Delaunay triangulation (green lines) and Voronoi tessellation (blue lines) of populations around the Mediterranean basin. (b) Genetic barriers were computed using 100 bootstrap distance matrices (computed via Principal Components Analysis and ten-fold cross-validation) and are indicated by red lines; thickness of the lines increases with the statistical significance of the respective barriers. Notice that strong genetic barriers separate the North African populations from the South European populations, as expected by the presence of the Mediterranean.

Figure S3a

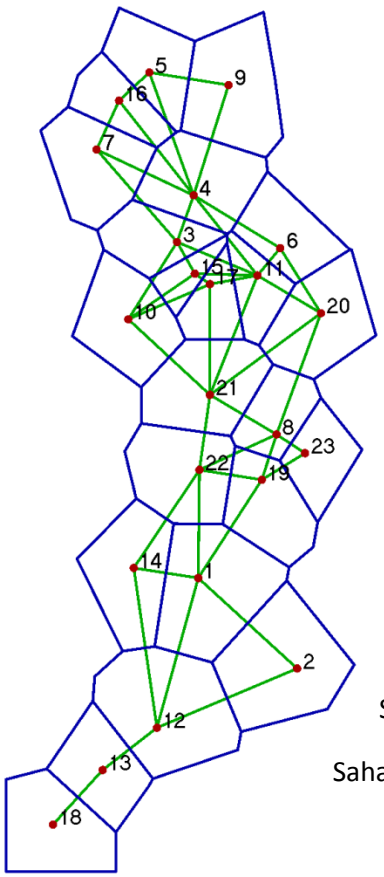
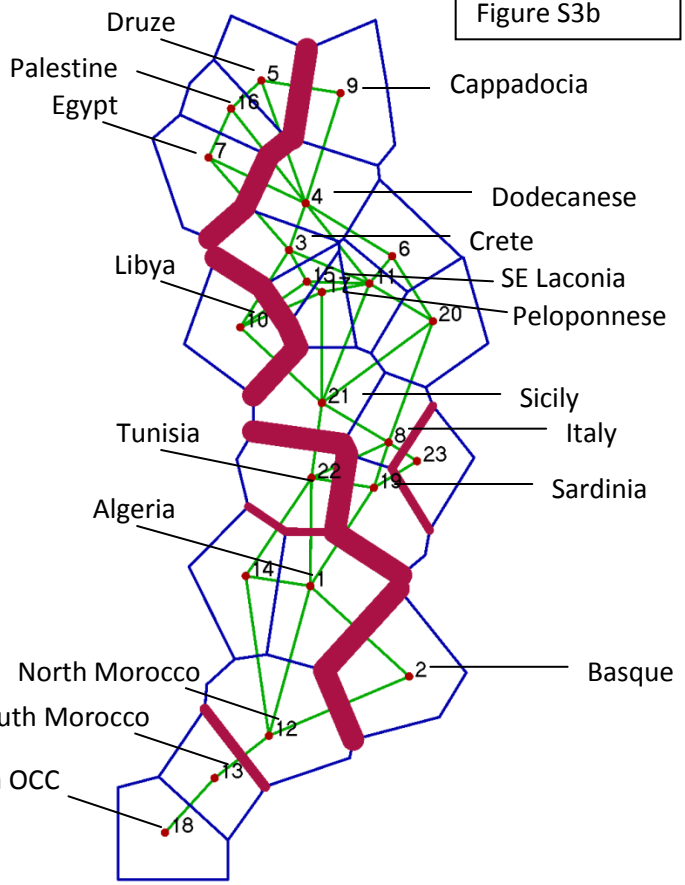
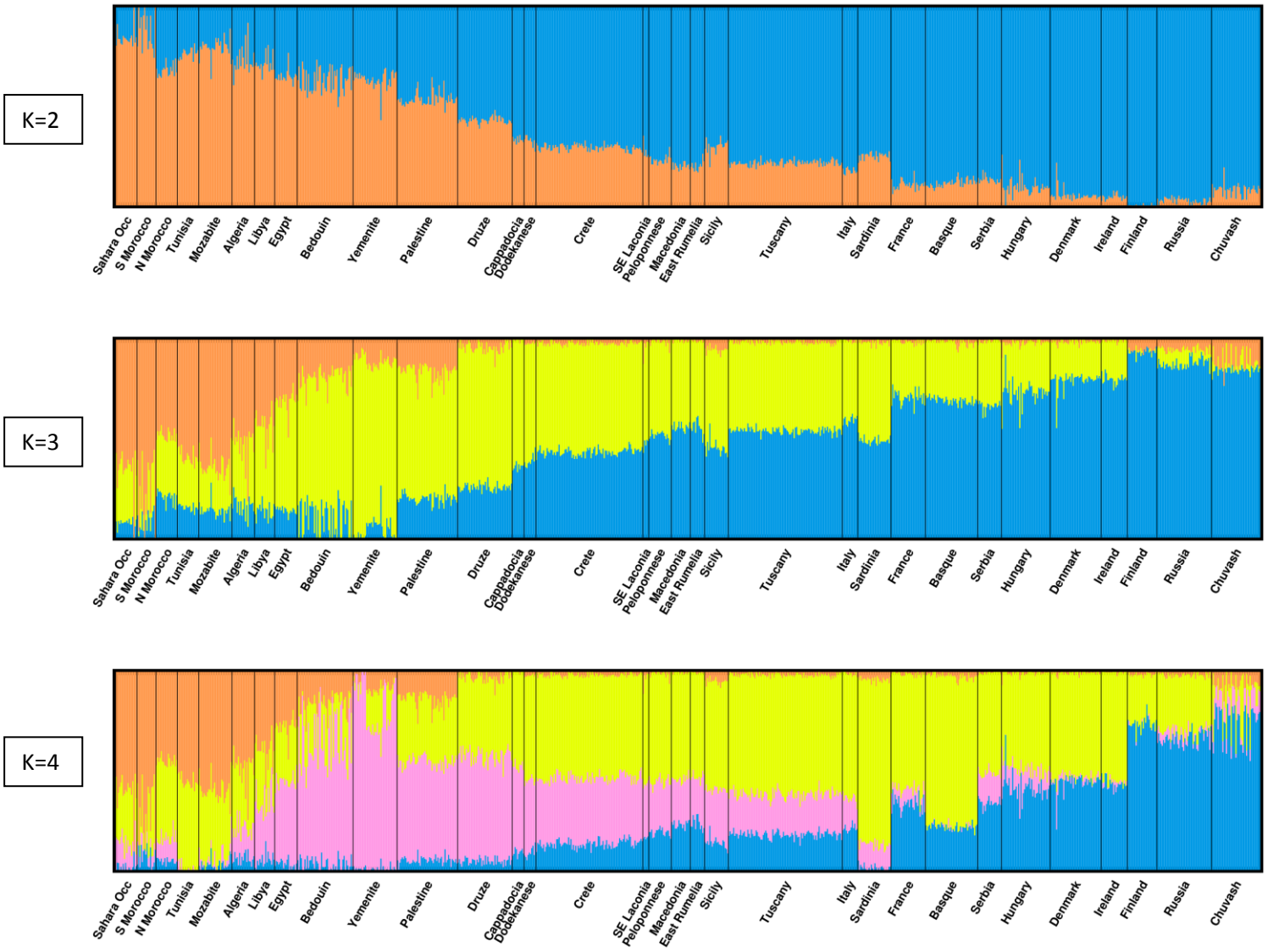


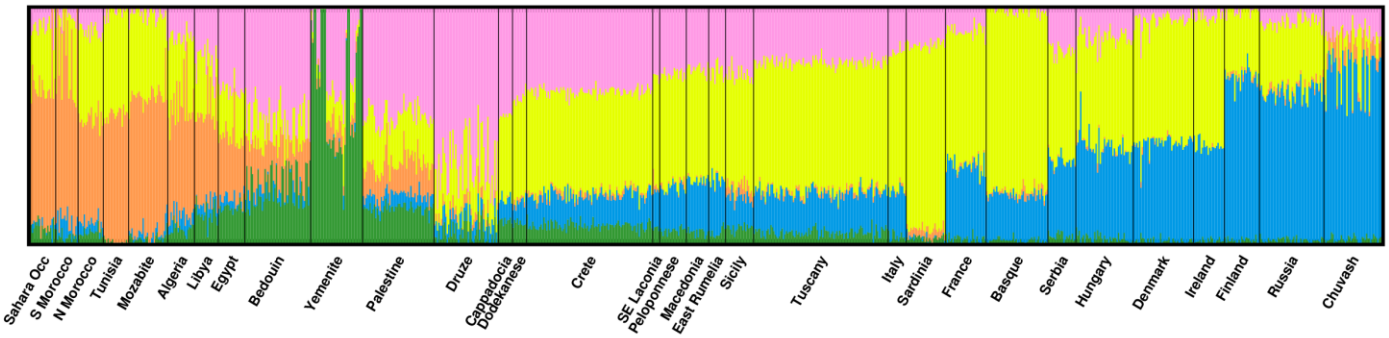
Figure S3b



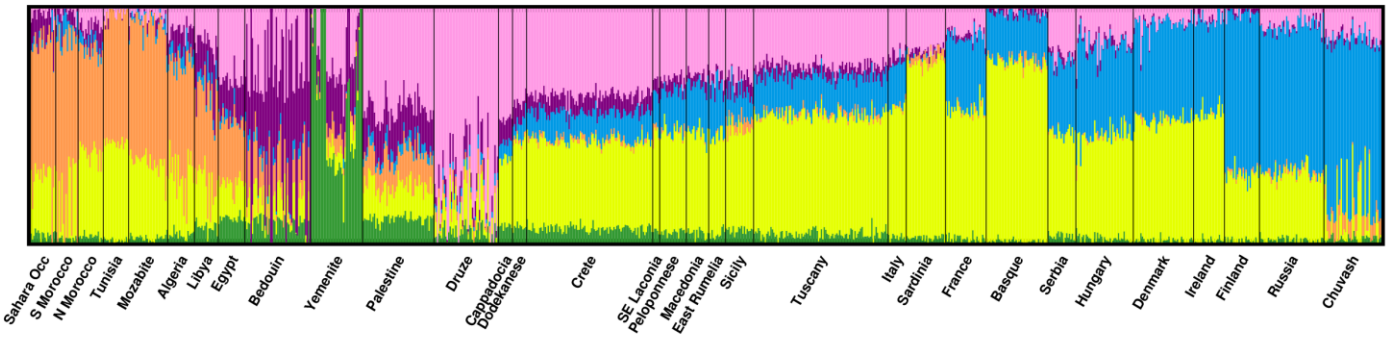
Supplementary Figure 4: ADMIXTURE plot of all populations in Supplementary Tables 1 and 2, for all values of K between two and eight. A clear gradient can be observed, from North Africa to Near East, South Europe (with a sub-gradient forming from Cappadocia, Dodecanese, Crete, and Peloponnesos) , and, eventually, North Europe.



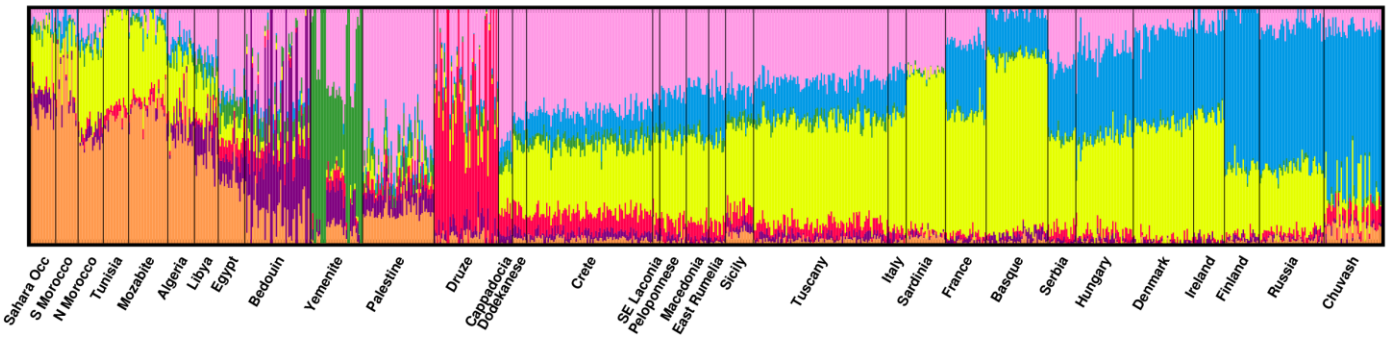
K=5


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Bedouin
Yemenite
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

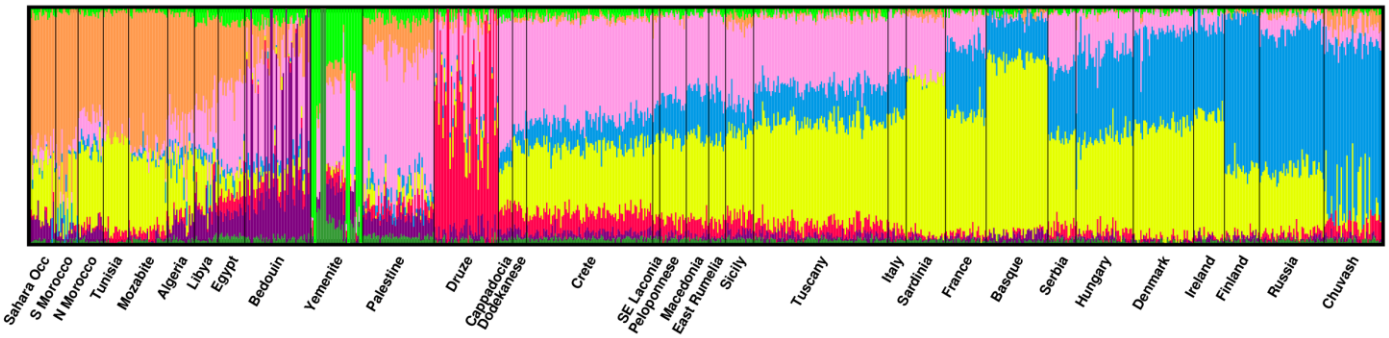
K=6


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Bedouin
Yemenite
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

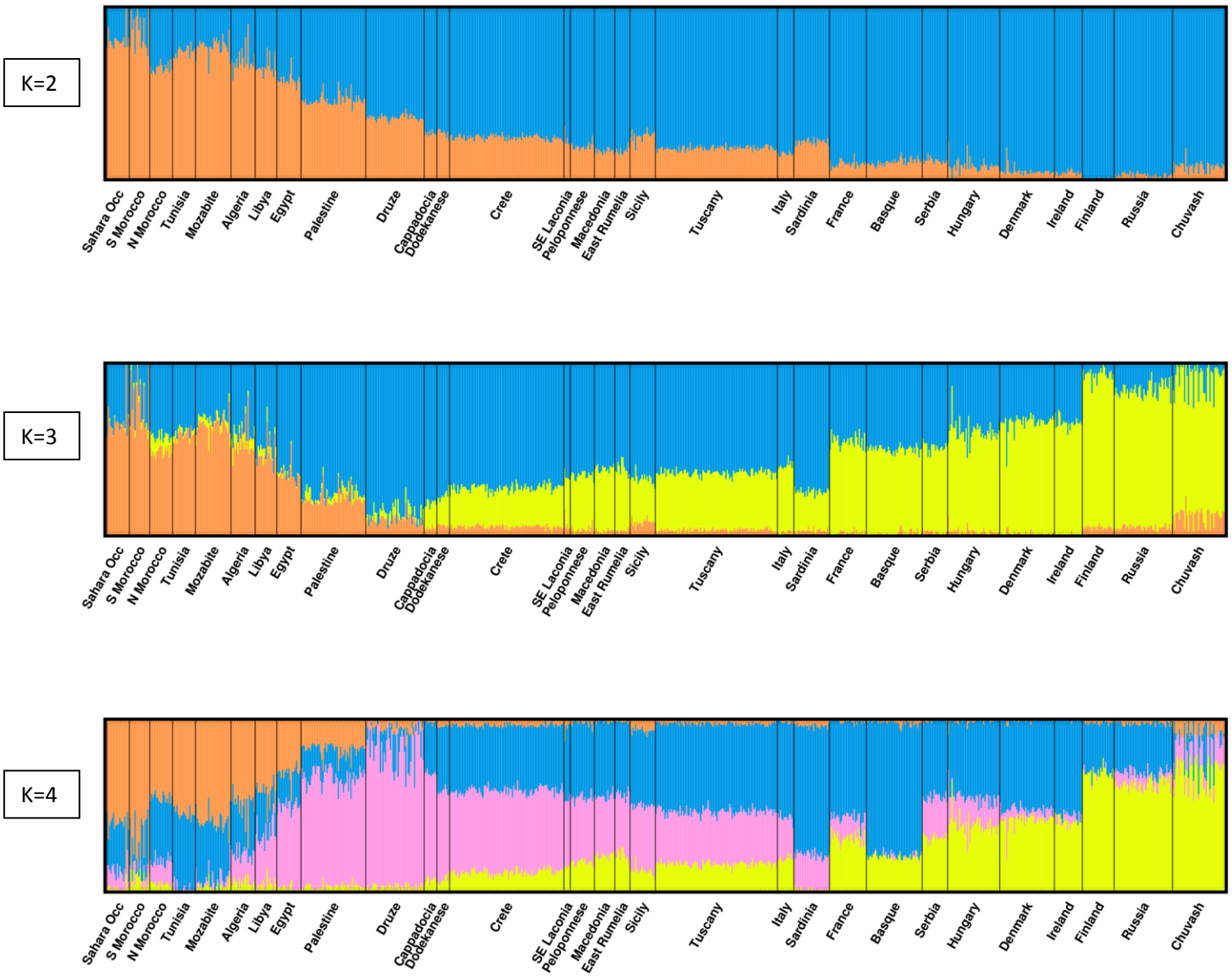
K=7


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Bedouin
Yemenite
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

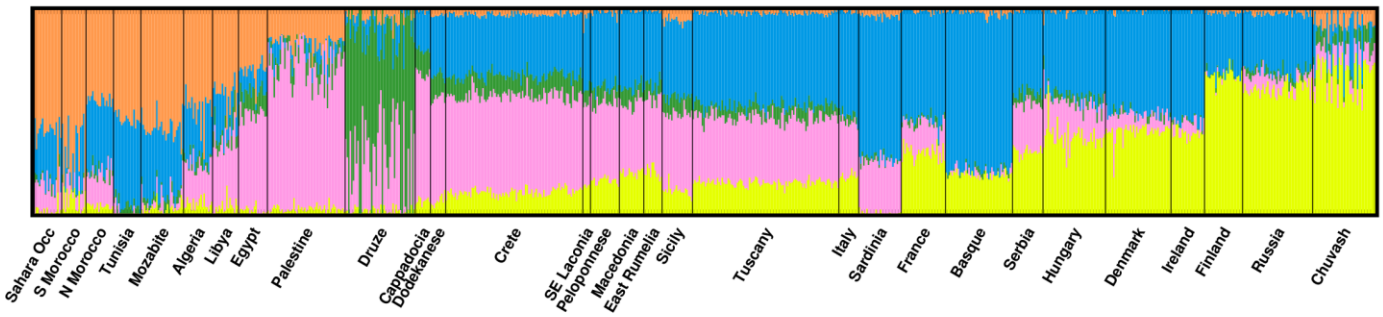
K=8


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Bedouin
Yemenite
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

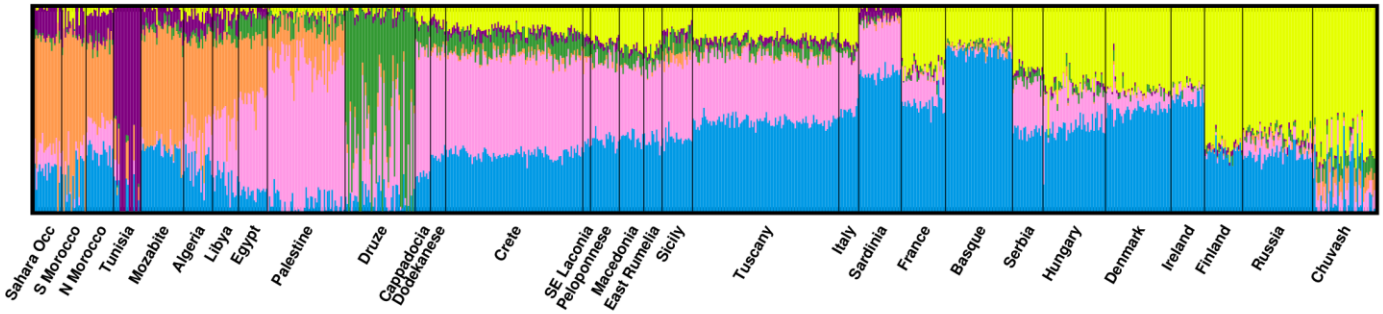
Supplementary Figure 5: ADMIXTURE plot of all populations in Supplementary Tables 1 and 2, excluding the Bedouins and the Yemenites, for all values of K between two and eight. We again observe a clear gradient from North Africa to Near East, South Europe (with a sub-gradient forming from Cappadocia, Dodecanese, Crete, and Peloponnesos), and, eventually, North Europe. The percentages of origin of each individual with respect to the K (unknown) ancestral populations were used in our network formation algorithms (see Supplementary Figure 8).



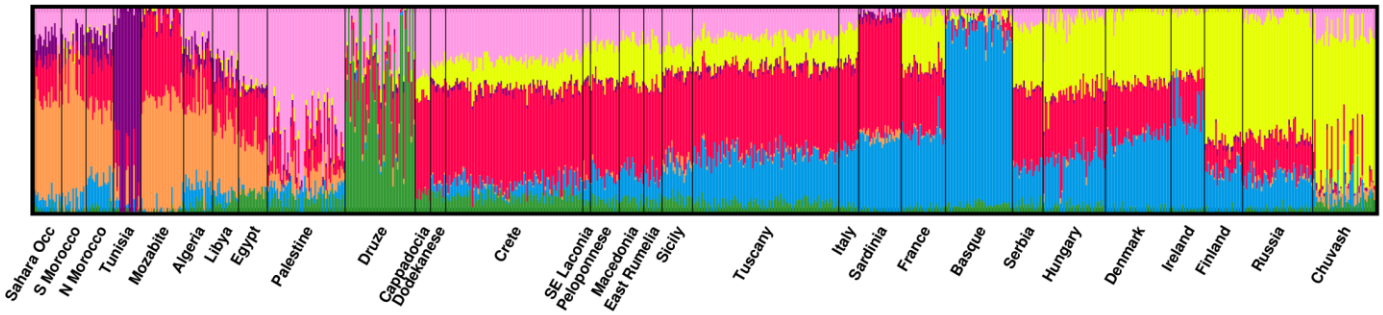
K=5


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

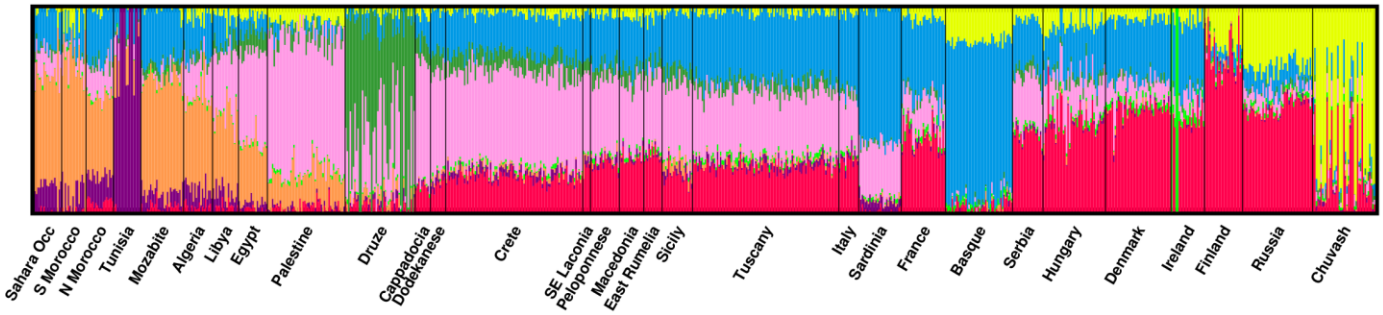
K=6


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

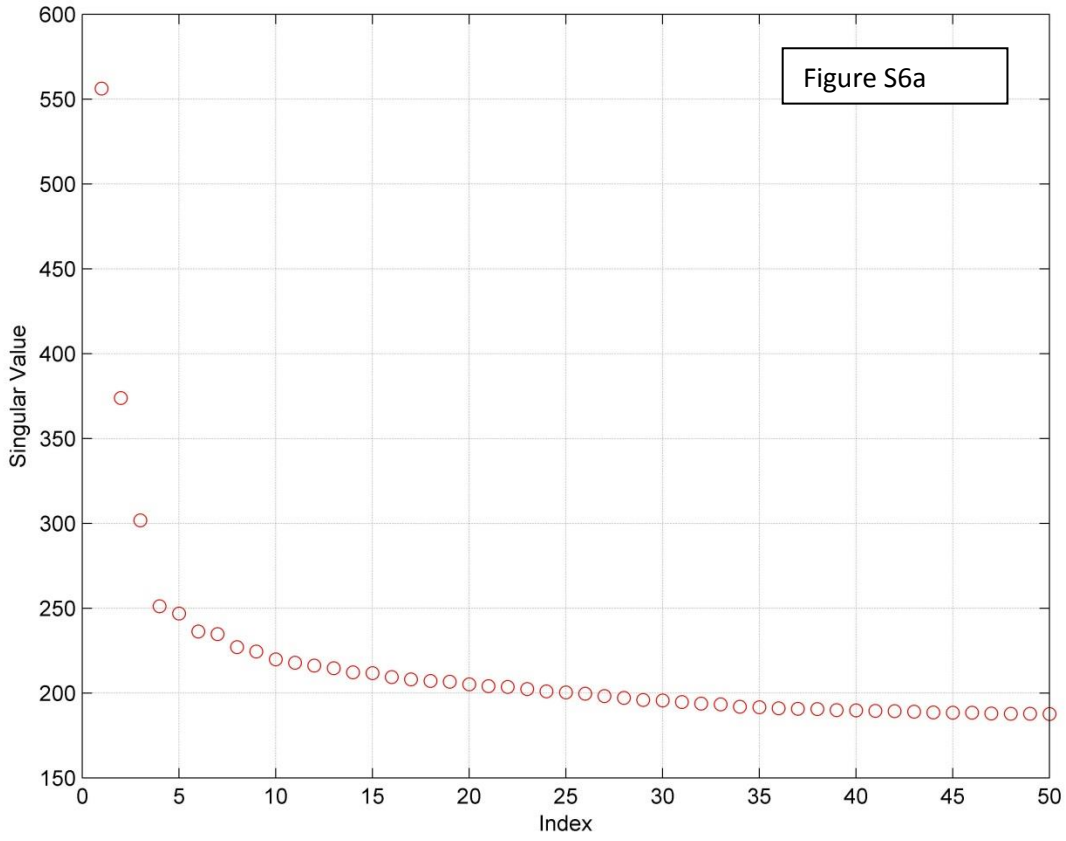
K=7


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

K=8


Sahara Occ
S Morocco
N Morocco
Tunisia
Mozabite
Algeria
Libya
Egypt
Palestine
Druze
Cappadocia
Dodekanese
Crete
SE Laconia
Peloponnese
Macedonia
East Rumelia
Sicily
Tuscany
Italy
Sardinia
France
Basque
Serbia
Hungary
Denmark
Ireland
Finland
Russia
Chuvash

Supplementary Figure 6: (a) Plot of top 50 singular values of the covariance matrix formed using all samples, excluding the Bedouins and the Yemenites, in Supplementary Tables 1 and 2. The top principal components of this covariance matrix will be used in our PCA-based network formation algorithms (Supplementary Figure 7). Applying a simple metric of singular value significance (i.e., extracting principal components corresponding to singular values that exceed the average singular value by at least three standard deviations) identifies the top seven singular values as significant. (b) Plot of top 50 singular values of the covariance matrix formed using all samples, including the Bedouins and the Yemenites, in Supplementary Tables 1 and 2. Applying a simple metric of singular value significance (i.e., extracting principal components corresponding to singular values that exceed the average singular value by at least three standard deviations) identifies the top ten singular values as significant.



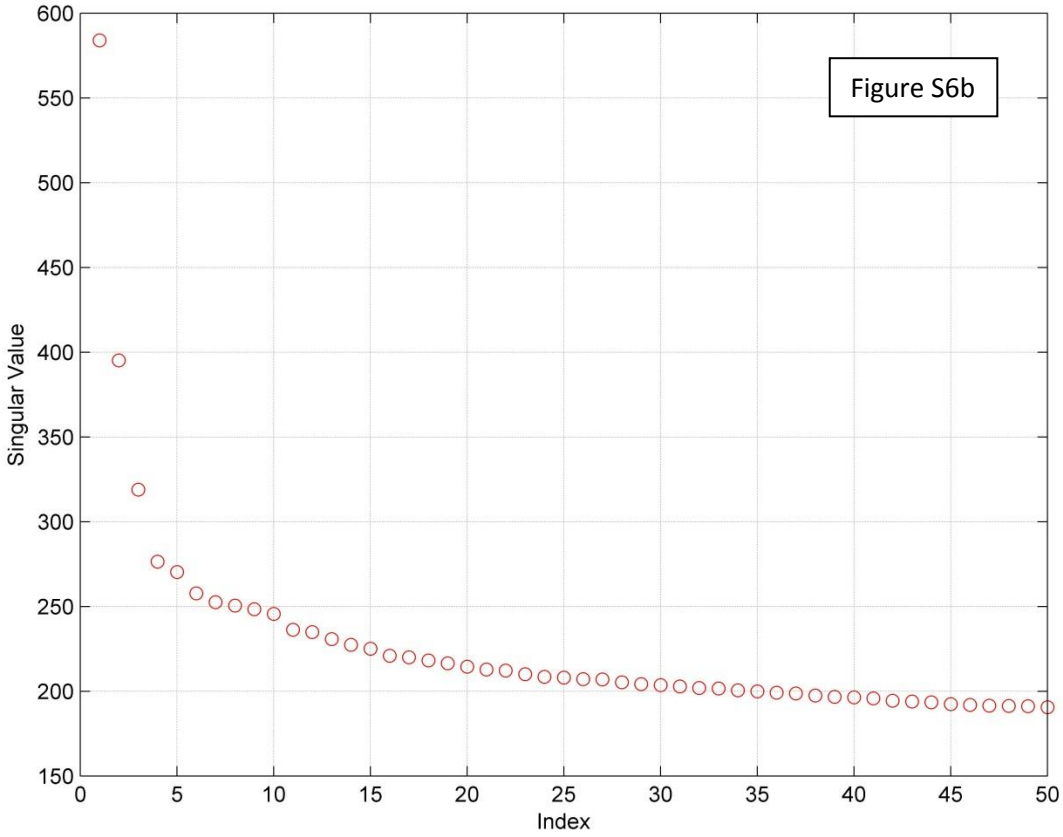
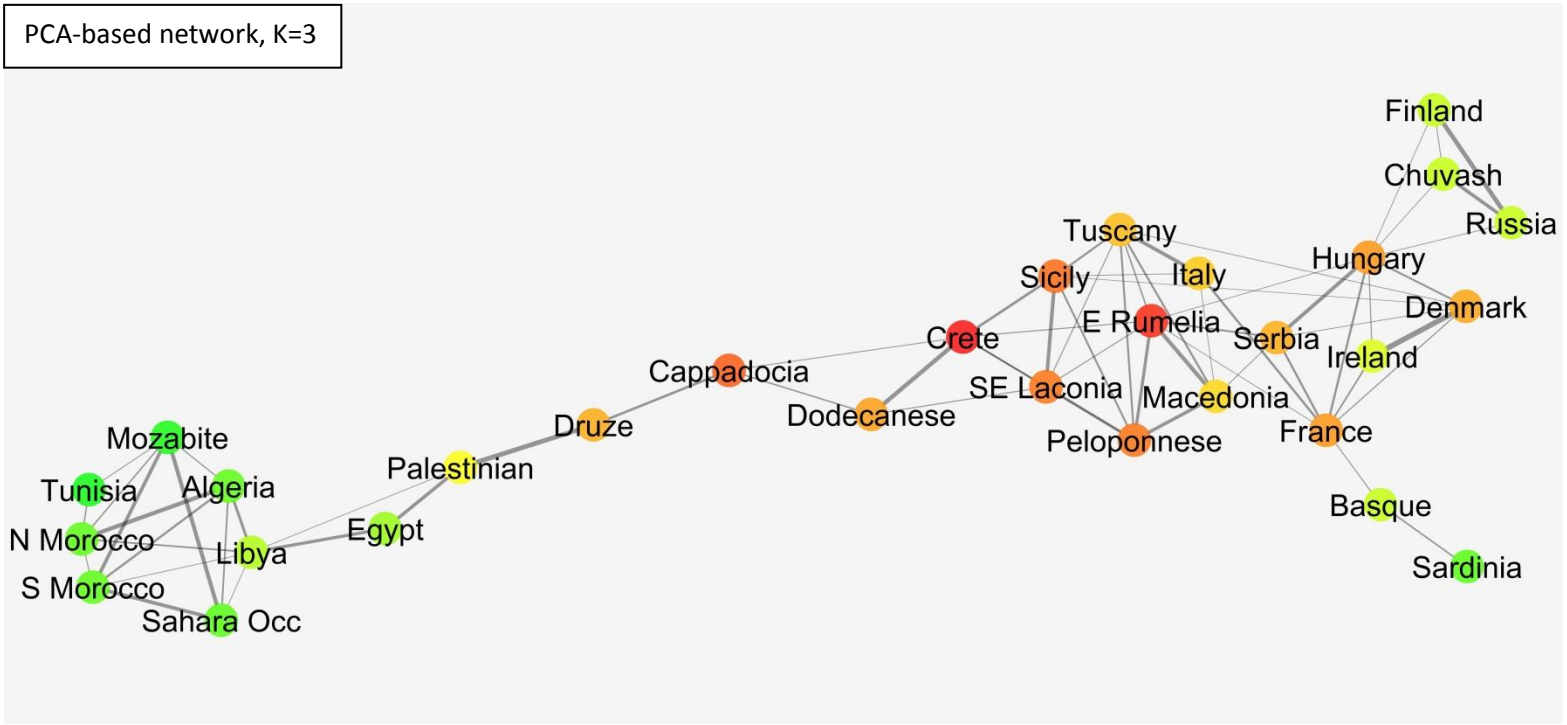
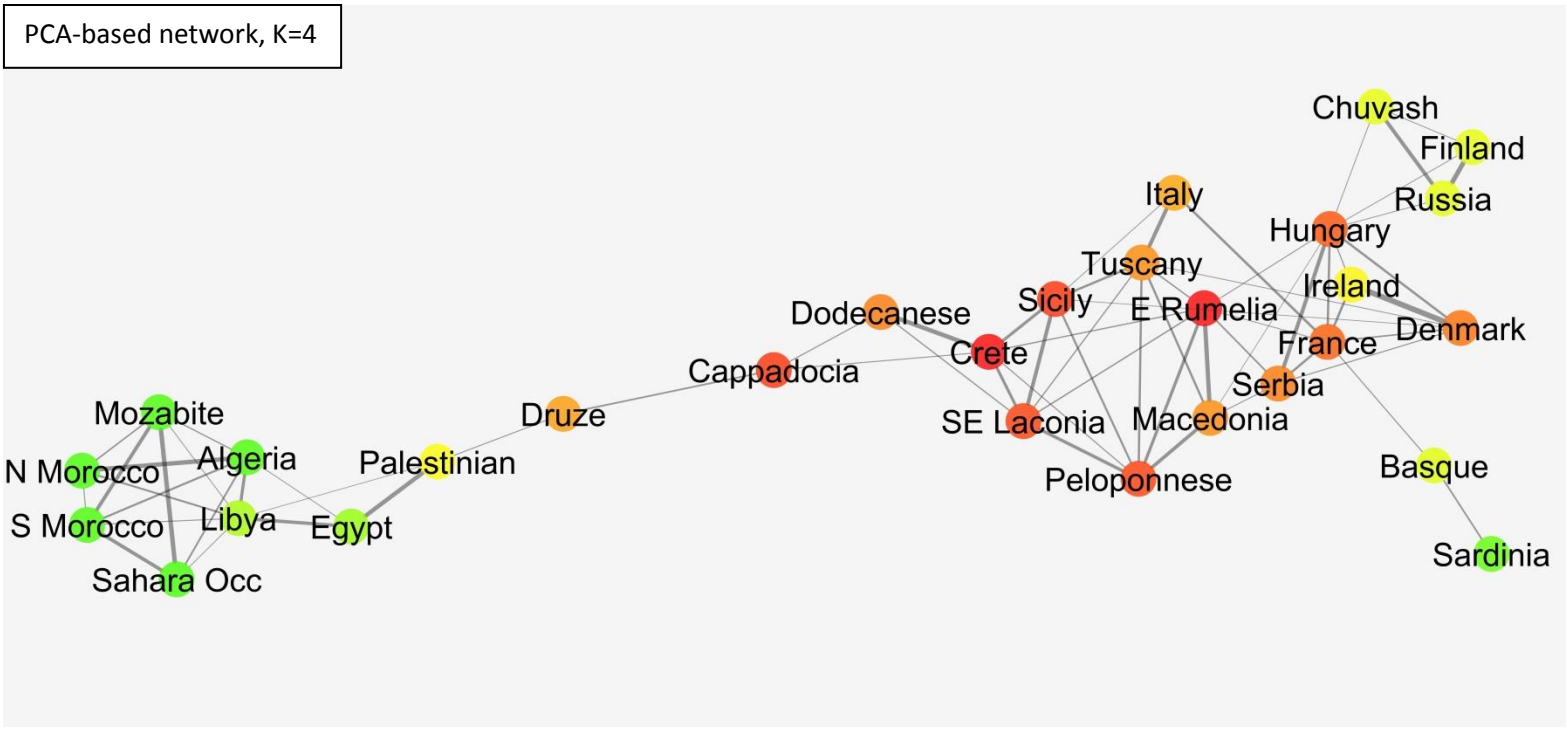


Figure S6b
Singular Value
Index
600
550
500
450
400
350
300
250
200
150
0
5
10
15
20
25
30
35
40
45
50

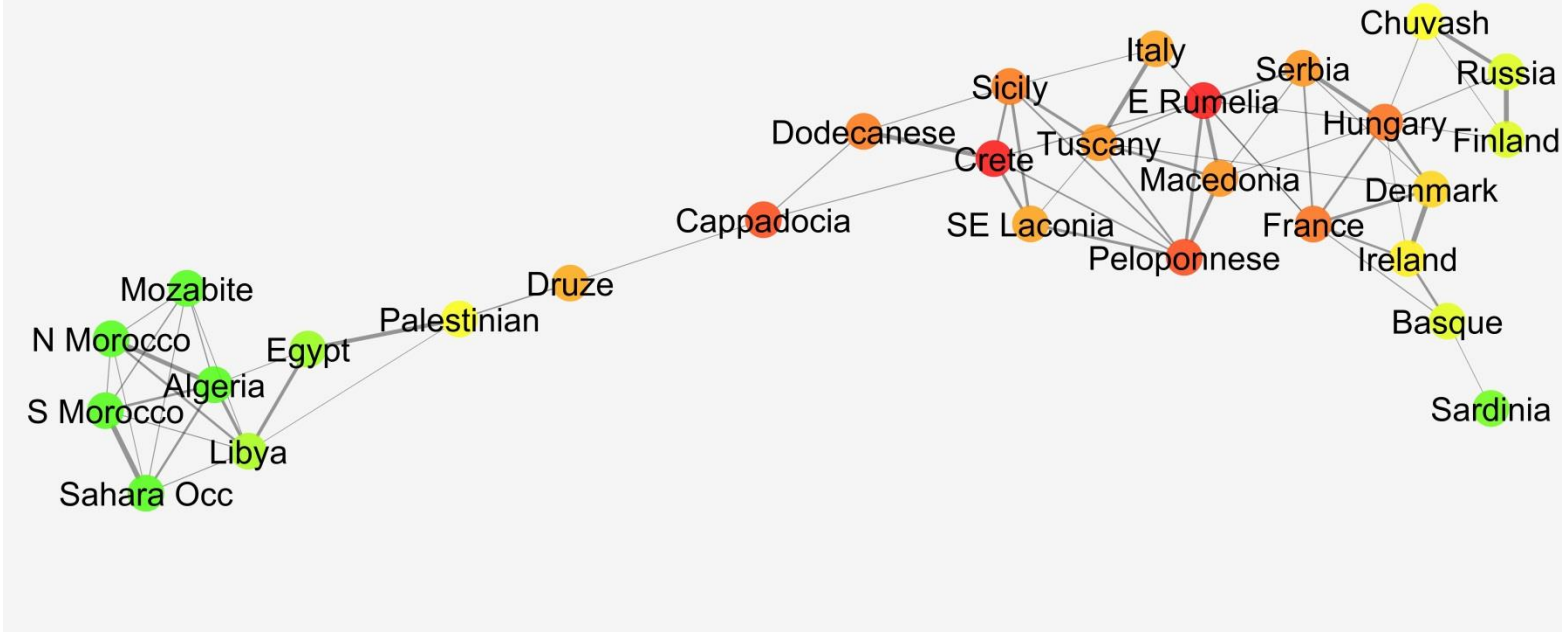
Supplementary Figure 7: Networks formed using PCA and the top K (for all values of K between three and seven) principal components in order to identify nearest neighbors for each individual (see Supplementary Methods for details and Supplementary Figure 6a for the number of significant Principal Components). "Warmer" colors indicate nodes of high centrality for the whole network (as computed by Cytoscape), while "thicker" edges indicate strong connections (high genetic similarity between the respective populations). Different values of K result in highly similar networks, highlighting the robustness of our results. The path from Northern Africa to Southern Europe via Near East, Cappadocia, the Dodecanese, and, of course, Crete is obvious. Note that the panel corresponding to K=5 is identical to Figure 4a in the main text; we include it here as well to facilitate comparisons.



PCA-based network, K=4

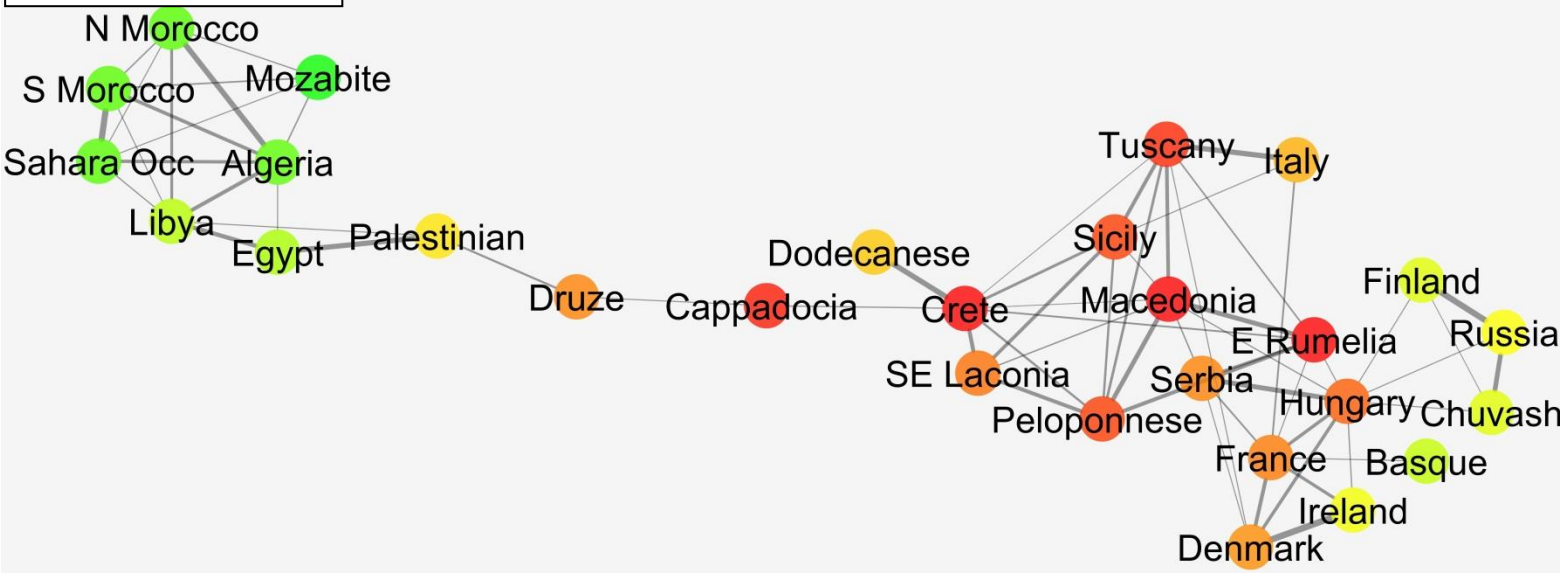


PCA-based network, K=5

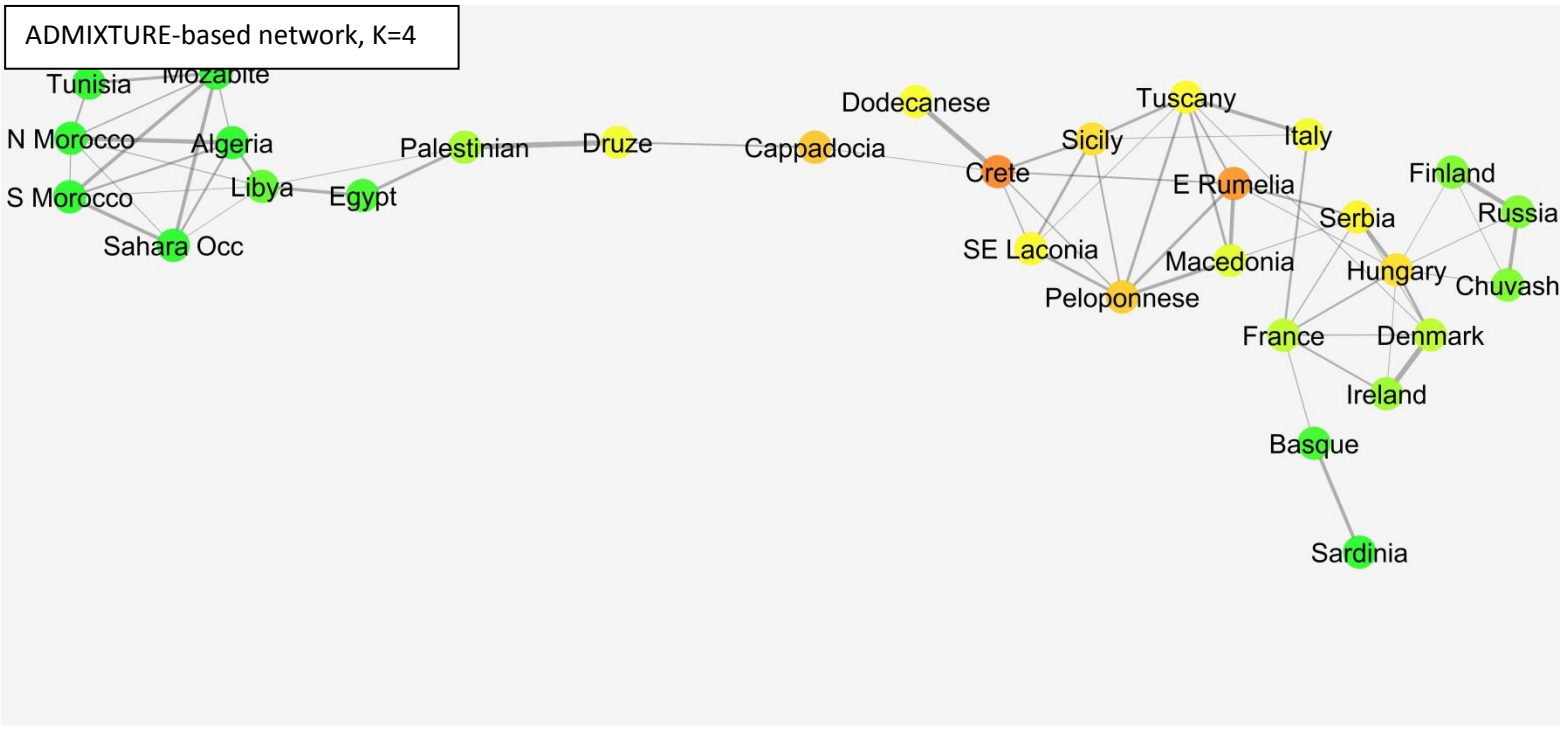


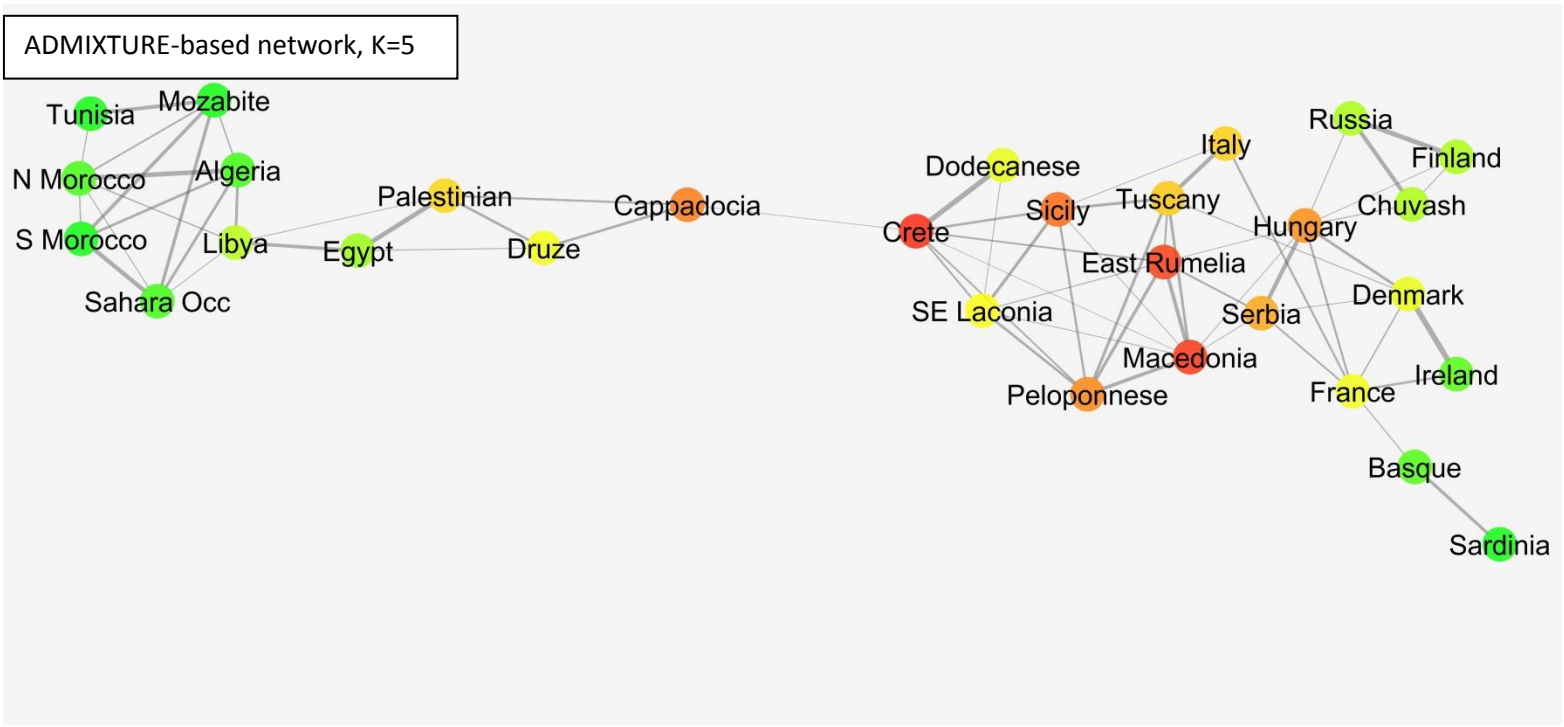
PCA-based network, K=6

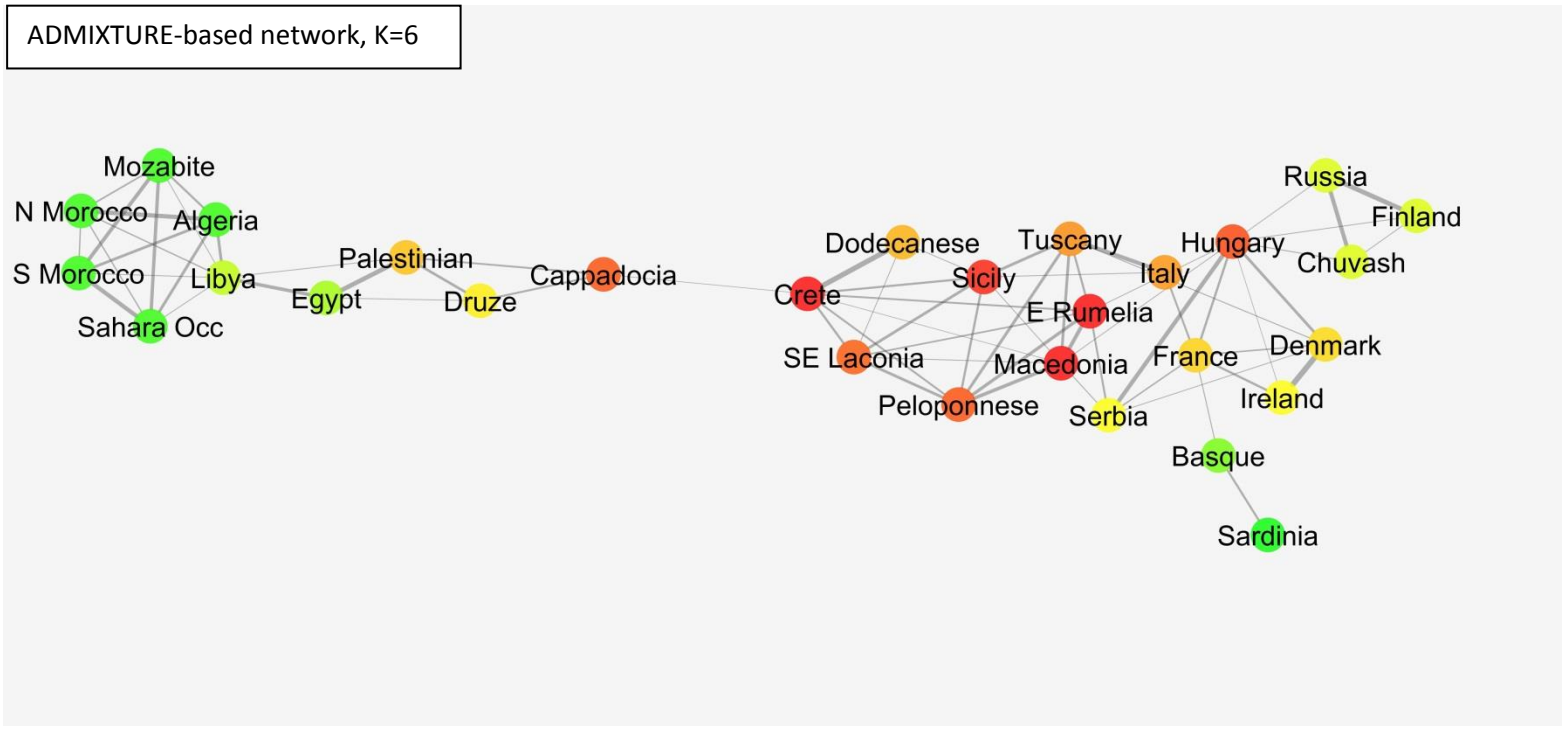
PCA-based network, K=7



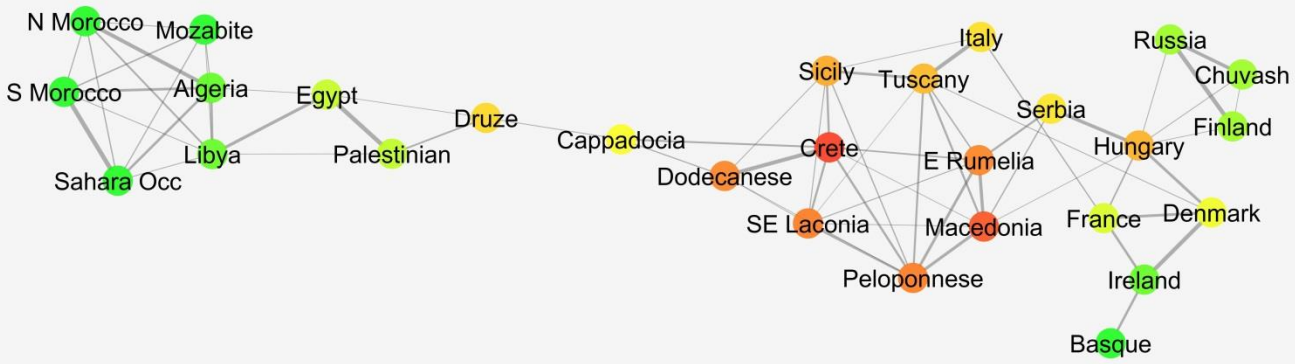
Supplementary Figure 8: Networks formed using ADMIXTURE with the parameter K (the number of ancestral populations) set between four and eight in order to identify nearest neighbors for each individual (see Supplementary Methods for details). "Warmer" colors indicate nodes of high centrality for the whole network (as computed and visualized by Cytoscape), while "thicker" edges indicate strong connections (high genetic similarity between the respective populations). Different values of K result in highly similar networks, highlighting the robustness of our results. Even more interestingly, the networks of Supplementary Figures 7 and 8 are again very similar, even though PCA is a model-free dimensionality reduction technique, while ADXMIXTURE is a model-based approach (see Supplementary Methods for a more detailed discussion of the two approaches). Once more, the path from Northern Africa to Southern Europe via Near East, Cappadocia, the Dodecanese, and, of course, Crete is obvious. Note that the panel corresponding to K=5 is identical to Figure 4b in the main text; we include it here as well to facilitate comparisons.



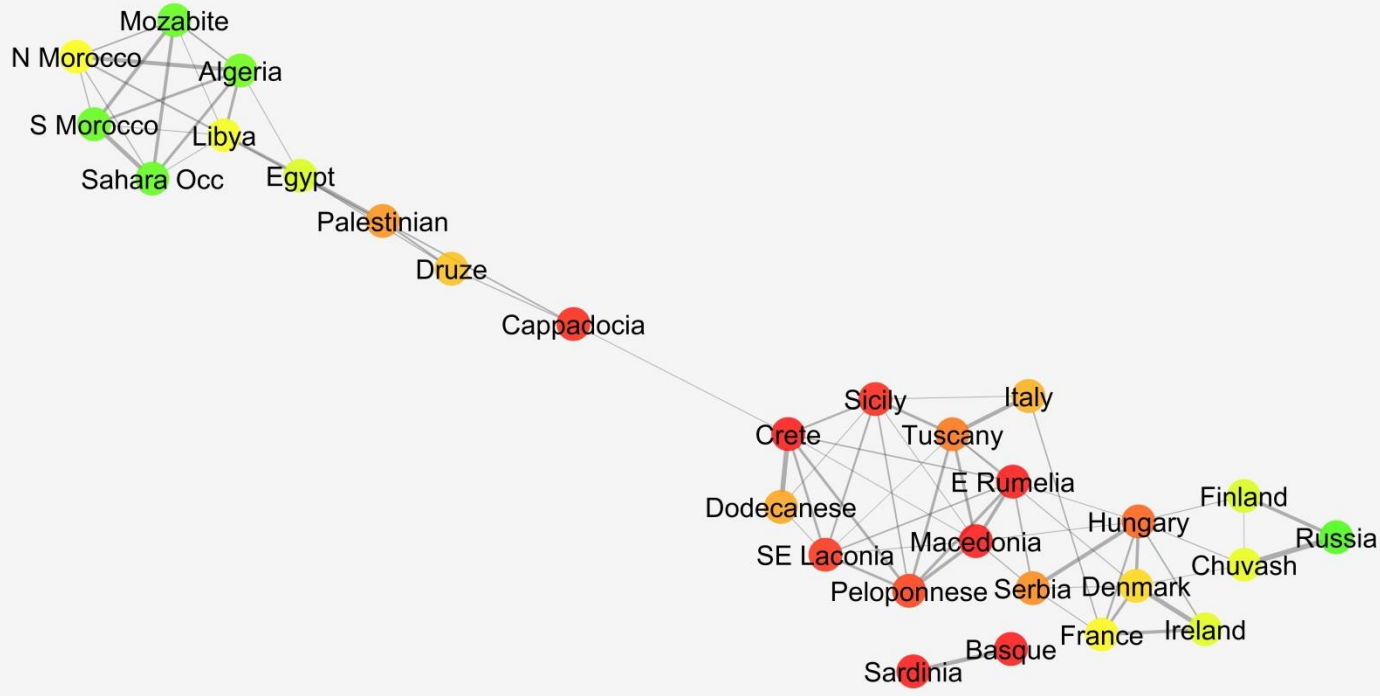

ADMIXTURE-based network, K=5
Tunisia
Mozabite
N Morocco
Algeria
S Morocco
Libya
Sahara Occ
Egypt
Palestinian
Druze
Cappadocia
Crete
Dodecanese
Sicily
Tuscany
Italy
Russia
Finland
Chuvash
Hungary
East Rumelia
SE Laconia
Serbia
Denmark
Macedonia
Peloponnese
France
Ireland
Basque
Sardinia


ADMIXTURE-based network, K=6
Mozabite
N Morocco
Algeria
S Morocco
Libya
Sahara Occ
Egypt
Palestinian
Druze
Cappadocia
Crete
Dodecanese
Sicily
Tuscany
Italy
Russia
Finland
Chuvash
Hungary
E Rumelia
SE Laconia
Macedonia
Peloponnese
Serbia
France
Denmark
Ireland
Basque
Sardinia

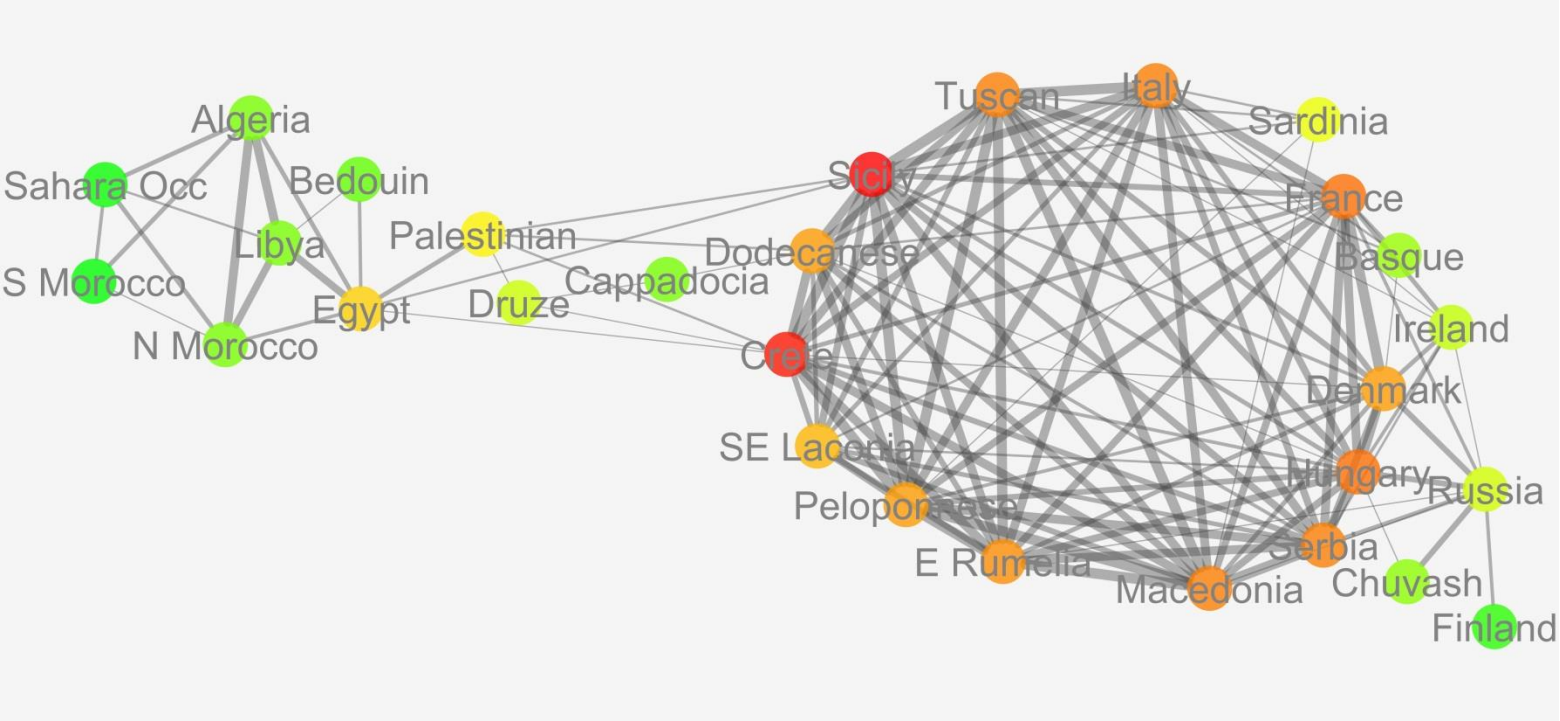
ADMIXTURE-based network, K=7



ADMIXTURE-based network, K=8



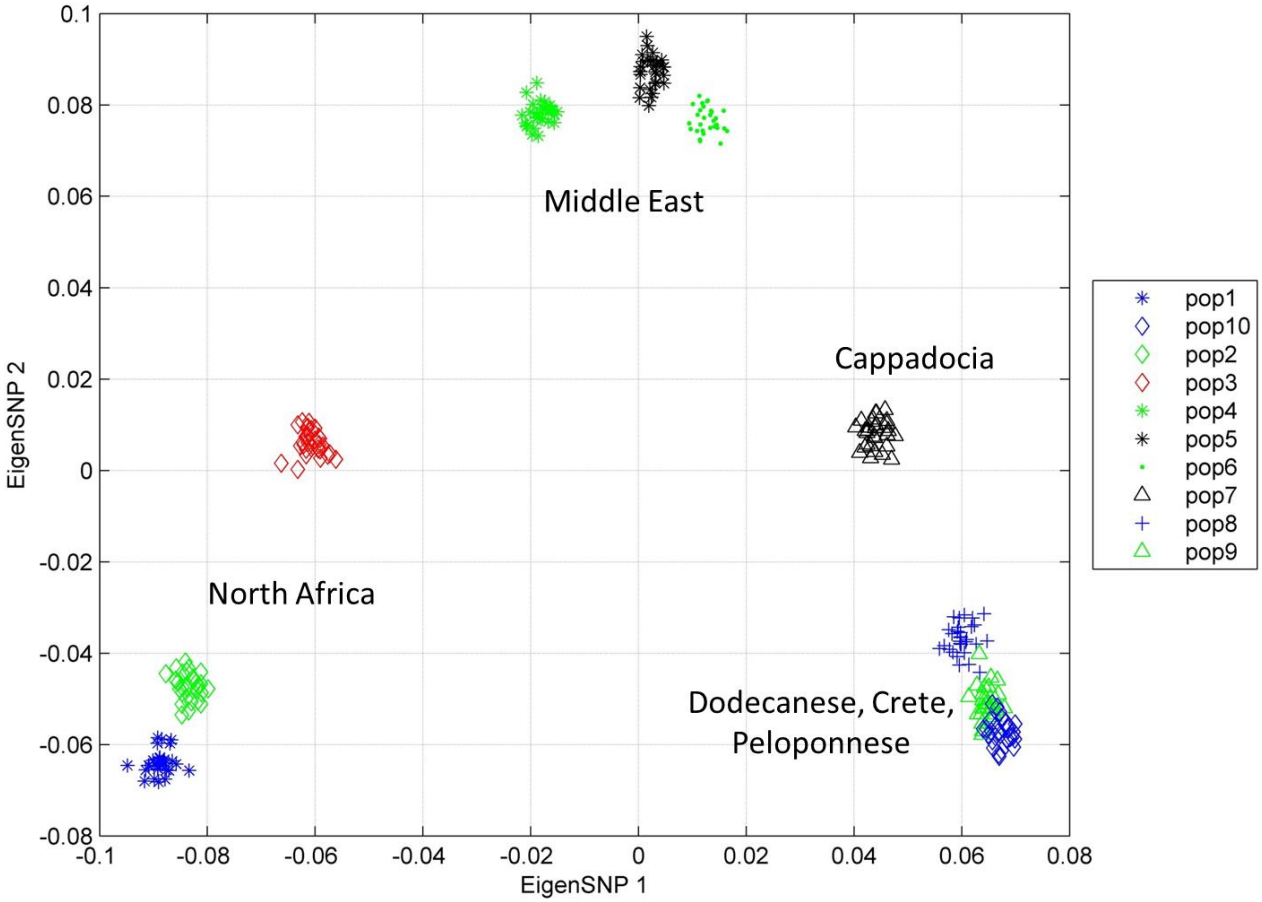
Supplementary Figure 9: Network formed using Fst values between pairs of populations (see Supplementary Methods for details). “Warmer” colors indicate nodes of high centrality for the whole network (as computed by Cytoscape), while “thicker” edges indicate strong connections (high genetic similarity between the respective populations).



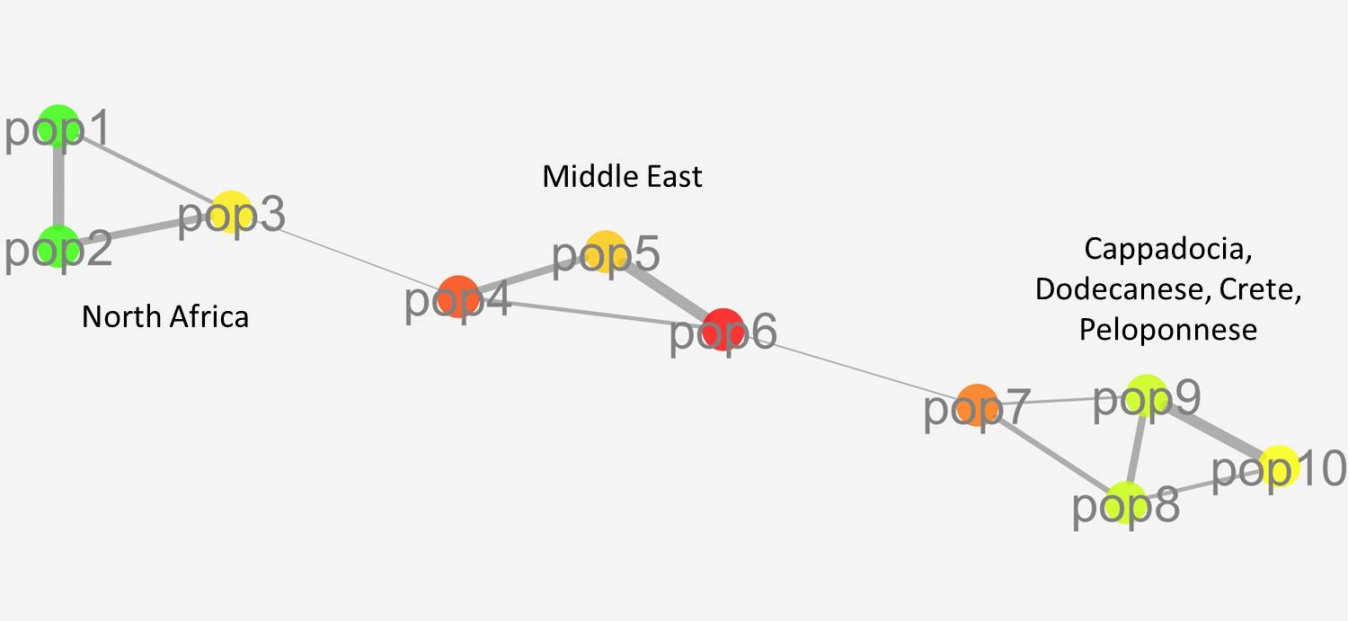
Supplementary Table 4: Geographic coordinates of the ten populations around the Mediterranean basin that were used as a basis for our simulations in Supplementary Figures 10 and 11. The table also indicates the distance from one population to the next one, the normalized distance from one population to the next one, and the respective lattice points where the populations were placed. We denoted the ten populations as **pop1** through **pop10** in Supplementary Figures 10 and 11.

Population	**Latitude**	**Longitude**	**Distance (in kms) from previous population**	**Normalized distance (divided by minimum distance)**	**Coordinates in x-axis (rounded)**
Algeria	36.74	3.03	0	1	1 **(pop1)**
Tunisia	36.82	10.15	633	3.0485	3 **(pop2)**
Libya	32.13	20.06	1046	6.4337	6 **(pop3)**
Egypt	30.04	31.23	1069	9.8932	10 **(pop4)**
Druze	33.51	36.31	634	11.945	12 **(pop5)**
Palestine	31.52	34.45	374	13.1553	13 **(pop6)**
Cappadocia	38.74	35.48	806	15.7638	16 **(pop7)**
Dodecanese	36.44	28.22	688	17.9903	18 **(pop8)**
Crete	35.34	25.14	309	18.9903	19 **(pop9)**
Peloponnese	37.51	22.37	344	20.1036	20 **(pop10)**

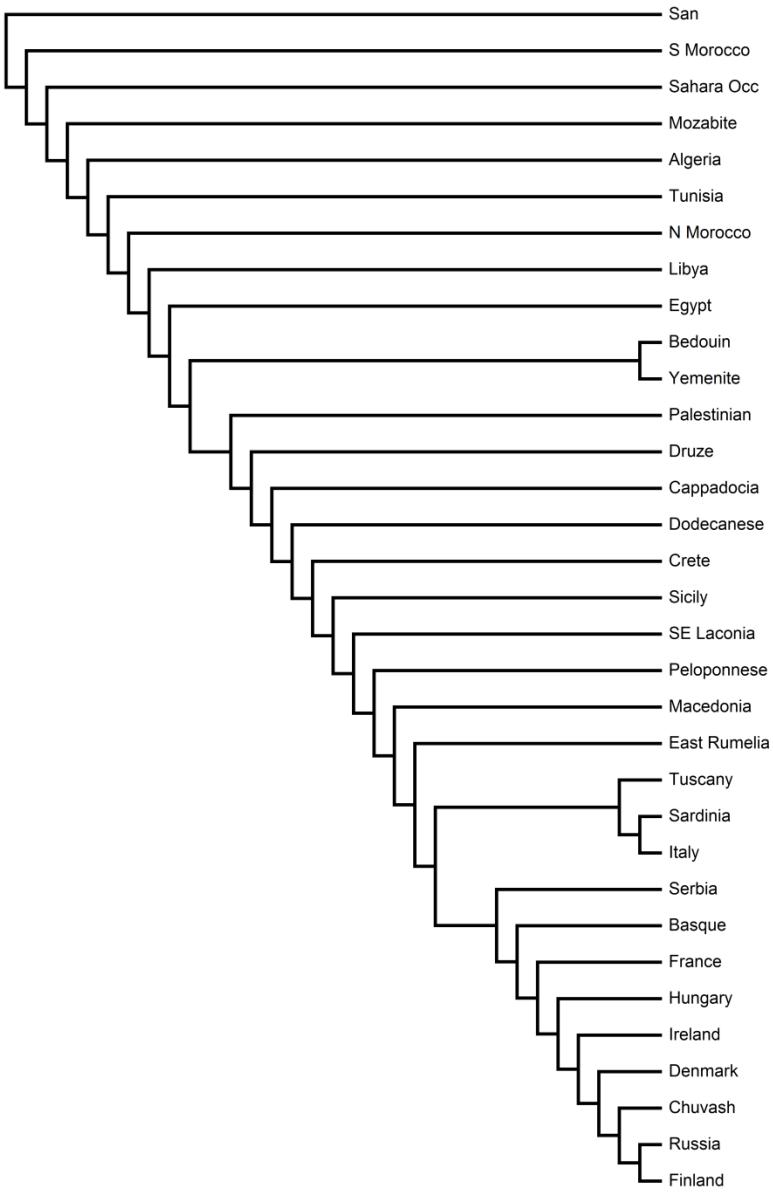
Supplementary Figure 10. PCA plot of ten simulated populations; see Supplementary Table 4 for a list of the corresponding populations around the Mediterannean basin that were used as a basis for our simulation. Note that the figure is reminiscent of our PCA plots of populations around the Mediterranean basin.



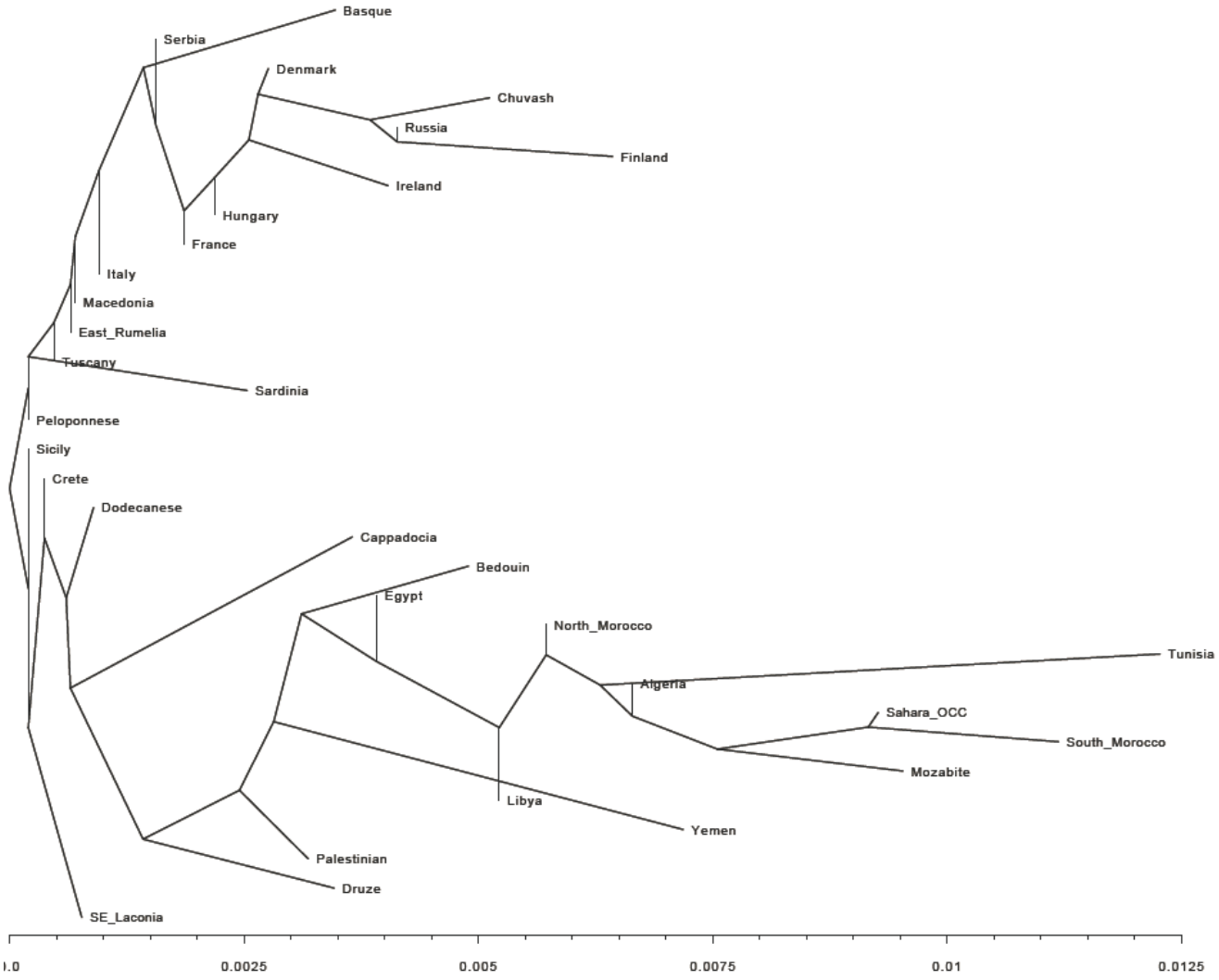
Supplementary Figure 11. Network formed by our algorithm using the PCA plot of data simulated under a stepping stone model (Supplementary Figure S10). In order to simulate the stepping stone model of migrations, distances between real populations around the Mediterranean were used (see Supplementary Table 4). Note that the network of the simulated data is in complete concordance with the network formed using real data for the respective populations.



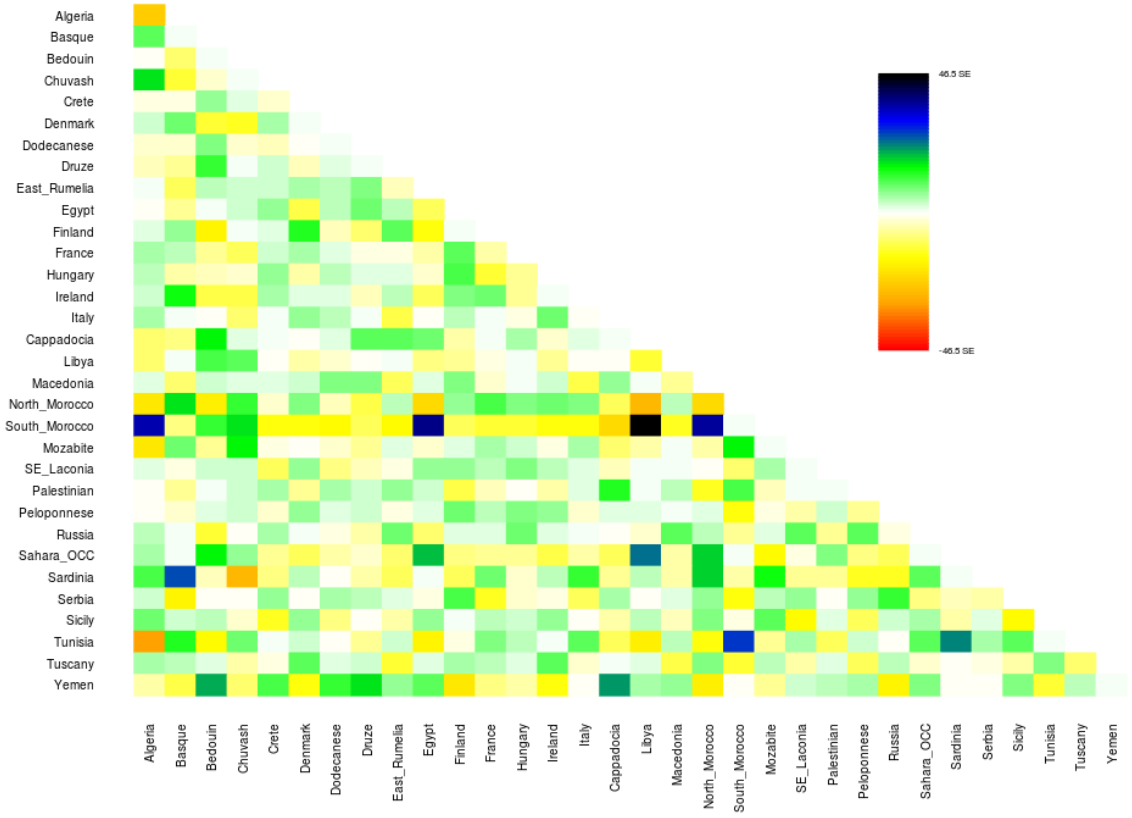
Supplementary Figure 12. Phylogenetic tree formed using Fst distances between all populations studied. The results are concordant with a closer genetic relationship of Anatolia to the islands of Crete and the Dodecanese rather than to the Balkans.



Supplementary Figure 13a. The maximum-likelihood tree generated by TreeMix, capturing 96.27% of the data. In concordance with all other analysis that we present, gene flow from Anatolia to Southern Europe appears to have occurred through the islands of the Dodecanese and Crete.



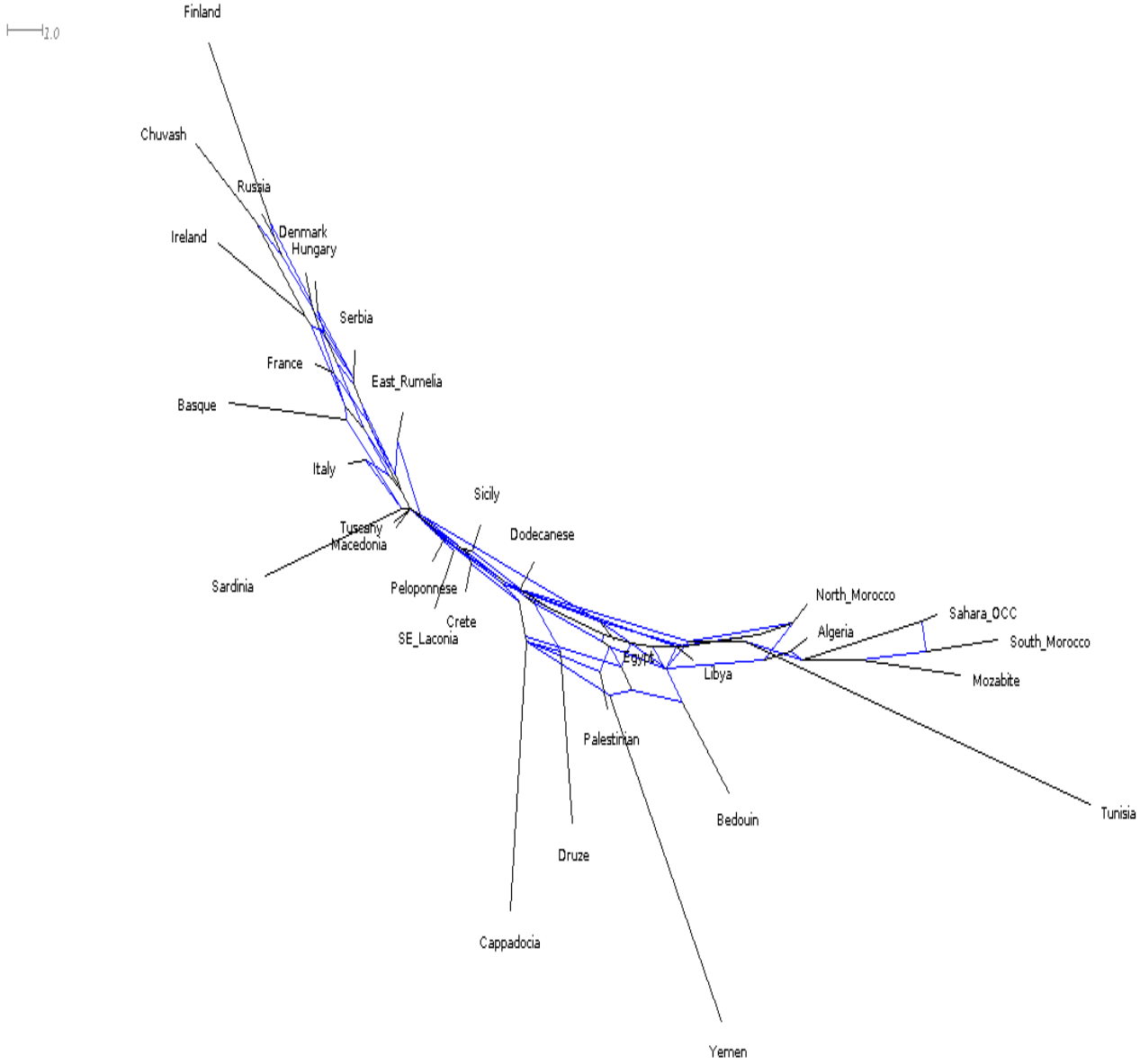
Supplementary Figure 13b. Residuals of the maximum-likelihood phylogenetic tree generated by TreeMix and shown in Supplementary Figure 13a. Population pairs that show high residuals (i.e., are not perfectly captured by the graph in Supplementary Figure S13a), were subsequently analyzed in order to identify pairs of populations that are more related to each other and corresponding migration events were inferred. Supplementary Table 5 shows the top ten inferred migration events based on this residual plot. As implemented in TreeMix, the residual covariance between each pair of populations *i* and *j* is divided by the average standard error across all pairs. This scaled residual in then plotted in each cell *(i,j)*. Colors are described in the palette on the right. Residuals above zero represent populations that are more closely related to each other in the data than in the best-fit tree and thus are candidates for admixture events.



Supplementary Table 5: Top 10 migration events (inferred by TreeMix), that improve the fit of the maximum likelihood tree shown in Supplementary Figure 13a. The respective migration weights are also shown. The original graph captured 96.27% of the available data, while the additional top 10 migrations inferred by TreeMix increase this percentage to 98.43%.

Migration weight	From	To
0.298771	South Morocco	Egypt
0.440767	South Morocco	Algeria
0.251526	Sardinia	North Morocco
0.251369	Sardinia	North Africa
0.05014	South Morocco	Chuvash
0.341009	Sardinia	Basque
0.114157	South Morocco	Palestinian
0.126861	South Morocco	Bedouin
0.481369	Bedouin	Libya
0.20762	Sardinia	Italy

Supplementary Figure 14. The phylogenetic network generated by the NeighborNet algorithm as implemented in the SplitsTree4 software package. As in all previous analyses that we presented, the islands of Dodecanese and Crete play a pivotal role in gene flow from Anatolia to Southern Europe.



Supplementary Material: Online Resources

(a) Raw data for all populations in Supplementary Table 1, collected under the auspices of this study, are available at:

http://www.cs.rpi.edu//~drinep/Maritime_Route/RAW_DATA/

(b) Cytoscape files for PCA and ADMIXTURE networks.

We have made available all PCA and ADMIXTURE networks as Cytoscape (.cys) files. Both files can be opened using the Cytoscape software (http://www.cytoscape.org/) and include numerous statistics regarding the networks (e.g., edge weights, path lengths, centralities, etc.).

The files are available at:

http://www.cs.rpi.edu//~drinep/Maritime_Route/NETWORKS_ADMIXTURE.cys

and

http://www.cs.rpi.edu//~drinep/Maritime_Route/NETWORKS_PCA.cys

(c) An Excel file with all pairwise Fst values (as computed by SmartPCA) for all pairs of populations in Supplementary Tables 1 and 2.

The file is available at:

http://www.cs.rpi.edu//~drinep/Maritime_Route/Fst.xls

(d) An Excel file with the f3 statistics for all triplets formed from the seven populations bridging Anatolia to Southern Europe (Cappadocia, Dodecanese, Crete, SE Laconia, Peloponnese, Macedonia, East Rumelia).

The file is available at:

http://www.cs.rpi.edu//~drinep/Maritime_Route/f3.xls