Supplementary Note 1:

Population Sampling

Non-Jewish samples: A recent whole genome genotyping of the entire HGDP-CEPH panel of 51 populations has yielded global coverage of genomic variation among modern human populations[1], but has provided only limited coverage of populations important for the aims of the present study. Therefore, in addition to the 14 novel Jewish communities, we also genotyped samples from a total of 27 non-Jewish populations, of which 26 are not represented at the HGDP-CEPH panel[1] (**Supplementary Table 1**). As samples from Andalusia and Catalonia are designated together as Spanish, we refer to the total number of newly collected populations throughout the text to be 25. Of these 25 newly studied populations, 17 are from non-Jewish "host" populations across the range of Jewish communities where Jews have resided during the past century (i.e., from the Caucasus, Europe, North and East Africa, South and Central Asia) (**Supplementary Table 1**). A set of 9 non-Jewish populations – Armenians, Cypriots, Egyptians, Jordanians, Lebanese, Saudis, Syrians, Turks, and Yemenis – that provide further information on the Levant and contiguous geographic regions, are included. The Cypriot samples were collected from the Republic of Cyprus. Unless specifically designated as Jewish - any named population origin indicates it to be non-Jewish.

Jewish samples: Diaspora communities from which Jewish samples were collected refer to communities where Jews have resided during the past century. This is the most commonly accepted usage among scholars of Jewish history[2-4]. Thus, we do not mean 'Diaspora' to imply a shared origin of Jewish communities in an ancient Middle Eastern or Levant source population, as this is one of the questions that we wished to address in the current study. Of note, Ashkenazi Jewry referred to herein denotes collection from the communities detailed in **Supplementary**

Table 1. Samples from the Bulgarian and Turkish communities were combined in the analysis and are referred to as Sephardi Jews.

Analysis was carried out on two partitioning of the datasets: the entire data set of populations included in this study (83), which represent the major range of the Old World populations, and a subset of West Eurasian populations (38), which also excluded Ethiopian[5] and Indian Jews[6] (**Supplementary Table 1**). As our analysis is based on the merging of a previously reported dataset[1] and our novel one, it is worth elaborating on the measures taken to avoid potential biases related to the effect of genotyping center or batch effects in sample preparation. Because the genetic clusters coinciding with phylogeographic groups are robust to combining our data with published data, and since the genetic clusters observed in the current study are concordant with previously published results[1,7,8], such batch effects do not seem to have significantly affected our findings. Furthermore, we grouped HGDP derived samples[1] and our own samples into separate clusters and carried out a Fisher's exact test of the randomness of the distribution of missing genotypes between the clusters. After obtaining p-values for each of the 226,839 SNPs used in downstream analyses, we estimated significance thresholds using a false discovery rate in the 'qvalue' R package. We used a threshold of 0.000019 at which 1% of tests are expected to be falsely rejected. In total 491 SNP markers were detected to have statistically significant differences ($p<0.000019$) in the proportion of missing genotypes between the HGDP dataset and our new samples. After removing these SNPs we performed additional QC and LD pruning steps in order to be sure that the results are robust. We kept SNPs above 99.9% genotyping success rate and excluded SNPs in linkage disequilibrium ($r2>0.2$) with nearby markers in a window of 1500 SNPs (advanced by 25 SNP). A total of 115,118 markers passed all these filtering steps and were used to repeat PC analysis for all PC plots presented in the study.

All the plots generated for comparison were indistinguishable from the plots presented in this study (data not shown).

Supplementary Note 2:

Details of Old-World PCA

We investigated the lower-ranked PC3 and PC4 for meaningful variation (**Supplementary Fig. 2b and 2c**) and describe each PC in terms of the two underlying genotype/allele frequency gradients, which can be inferred when SNPs loading the PC are extracted and their frequencies are estimated within population samples. When the graphs of allele frequencies are superimposed on the PC plot (data not shown), two opposing gradients are evident, each having the highest frequencies on the two extremes of the PC plot. Allele frequency gradients beginning on each side decline towards the zero point of the plot, where they overlap or disappear. We used this general methodology to describe coordinates of populations on the PC plots. Indeed, the analysis of the Old World dataset suggests that PCs beyond the top two add some extra information relevant to the subject of our study.

PC3 of the Old World dataset (**Supplementary Fig. 2b**) reflects two gradients of genotype variation, one begins in the group of Yoruba-Mandenka-Bantu and declines towards a group of African hunter-gatherers (Biaka Pygmy, San, Mbuti Pygmy) while the second gradient has an opposite direction and starts in African hunter-gatherers (Mbuti Pygmy). Thus, PC3 describes genotype variation within Africa and projection of Middle Eastern populations along this PC suggests that alleles in Yoruba-Mandenka-Bantu group are also present among a subset of Bedouins, but their frequency otherwise declines within the Middle East. It should be noted that coordinates for the majority of Eurasian populations along this PC are near-zero, indicating that alleles determining both parts of this axis have either zero or trace frequencies. In the context of our study, PC3 adds additional detail about the genetic structure of Ethiopian and

Yemenite Jews. Their coordinates suggest the presence of Mandenka, Yoruba and Bantu associated alleles, but not alleles present in African hunter-gatherers. Coordinates along this PC, in addition, also differentiate Yemenite Jews from the majority of Jewish Diaspora samples, which are clustered near zero like most Eurasian populations.

PC4 of the Old World dataset (**Supplementary Fig. 2c**) is particularly informative for clarifying the genetic affinities of the two Jewish groups residing in India[6,9]. PC2 coordinates (West-East Eurasian axis of genotype variation in **Supplementary Fig. 2a**) cluster the Indian Jewish communities with South Asian host populations. Thus, PC4 yields additional information, enabling distinction of South Asian from East Eurasian influence. The left side of PC4 describes genotype variation running from the Indian subcontinent (Southern Indian samples on the left extreme of the **Supplementary Fig. 2c**) towards the rest of Eurasia, where it declines and disappears. Therefore, projection of the two Jewish groups (Bene Israel and Cochin) near Indian and Pakistani populations supports the presence of a South Asian genetic influence. By way of comparison this South Asian contribution delineated by PC4 coordinates, is similar to that observed in Pakistani populations, but clearly much less than in South Asian host populations. It is also worth noting that PC4 coordinates are consistent with inferences from the ADMIXTURE analysis in the current study (**Fig. 3**, K=8 and **Supplementary Fig. 4a,** bar plot for K=8). The order of populations along PC4 moving from the left to right extremes is: South Indian, Pakistani, Iranian, Lezgin, and Central Asians. This correlates with the ADMIXTURE results presented in **Fig. 3**, K=8, where inferred membership in the dark-green component declines from Southern Indian toward Pakistani, Iranian, Lezgin, and Central Asian populations.

While sample sizes are too small to allow firm generalizations regarding the Belmonte and Uzbekistani (Bukharan) Jewish communities, assignment of the few individuals from these

communities in our survey is noteworthy (**Fig. 1** and **2** and **Supplementary Fig. 2** and **3**). Importantly, these samples are collected from individuals who self-identified as being affiliated with the original members of the community and are not influenced by very recent historical events such as the migration of Ashkenazi Jews within the former Soviet Union[10] (i.e., in the case of the Bukharan samples). The tiny Belmonte Jewish community, currently comprising a few hundred individuals[11], has maintained a crypto-Jewish lifestyle and tradition following their flight to Portugal following the Spanish expulsion and until their re-conversion to Judaism in recent decades. The single Belmonte Jewish sample included herein falls in the very center of the major Jewish cluster on PCA (**Fig. 1 and Supplementary Fig. 2a**), and is distant from the sampled Andalusian and Catalonian Iberians. The contemporary Bukhara Jewish community, currently comprising 150,000-200,000 individuals, maintains an oral tradition of ancestry attributed to the 8th century BCE exile of the Northern Kingdom of Israel by the Assyrians[12]. Considered one of the oldest ethno-religious groups of Central Asia, Bukhara Jews preserved their Jewish identity and heritage despite being cut off from the rest of the Jewish world for more than two millennia. Our analysis assigned one Bukhara sample to the dominant Jewish cluster, and the other was in close proximity, and far from Central Asian populations (e.g., Uzbeks) (**Fig. 1 and Supplementary Fig. 2a**).

Details of West Eurasia restricted PCA

Restricting the PCA to West Eurasian populations allows focusing on regional scale axes of variation and better delineation of genetic distances within this region (**Supplementary Table 1** and **Supplementary Fig. 3**). In order to clarify the resulting conglomerate of individuals on the plot we estimated population-wise kernel densities for samples with at least 10 individuals (**Fig. 2**). We used the adehabitat R package[13] to estimate "utilization distribution" of a PC plot space

by individuals (points with coordinates PC1 and PC2) from different populations. The contour of the 92% distribution range of each population was then obtained and plotted. A smoothing parameter was chosen using a least-square cross-validation method[14]. Obtained contours were superimposed on the PC plot for populations with less than 10 individuals.

The Old World based PCA identified a major Jewish cluster, comprised of Ashkenazi Azerbaijani-Georgian-Iranian-Iraqi-Moroccan-Sephardi-Uzbekistani Jewish communities and a second Yemenite Jewish cluster, both of which localized to within the boundaries of other contemporary Levantine populations (**Fig. 1** and **Supplementary Fig. 2a**). The major Levantine Jewish cluster obtained on the background of Old World populations PCA is further divided into two when the PCA is restricted to the set of West Eurasian populations, while the Yemenite Jewish cluster remains independent. These three clusters (**Supplementary Fig. 3**), comprised of Ashkenazi-Moroccan-Sephardi, Azerbaijani-Georgian-Iranian-Iraqi-Uzbekistani, and Yemenite Jewish communities might attest to pre-existing structure among a shared ancient Hebrew population, splintering early in the history of the Jewish Diaspora, or to gene introgression during the period of the Diaspora. With analysis restricted to West Eurasia, PCA splits the Eastern and Western Jewish communities, consistent with a correlation of genetic with geographic proximity. It is important to note that similar separation to the one observed using the restricted set of West Eurasian populations is also evident in the PC1/PC4 plot of the entire sample set of Old World populations (**Supplementary Fig. 2c**).

The positioning of the Ashkenazi-Moroccan-Sephardi cluster between contemporary European and Levantine populations is of interest. This intermediate location is not surprising for Ashkenazi Jews, and might intuitively suggest some gene introgression from Central and East European host populations[15]. However, the overlapping location of Moroccan and Sephardi

Jewish communities should be considered in the context of their historical chronicles. The traditional scenario suggested by historians to explain the establishment of these latter communities is as follows[3,4,10,12,16]: 1) migration from the Levant to the geographic region corresponding to contemporary Iraq some 2,500 years ago, 2) movement to North Africa during the Arab expansion beginning in the 7th century, 3) movement through Gibraltar to the Iberian Peninsula to merge with and probably overwhelm any small pre-existing Jewish community from the Roman era, 4) expulsion of an estimated greater than one third of the population from the Iberian Peninsula in 1492 (Spain) and 1497 (Portugal) ACE, and 5) re-settlement throughout the then existing North African Jewish communities or establishment of new Jewish communities such as the Turkish and Bulgarian communities under Ottoman rule[17]. Therefore, the clustering of Sephardi and Moroccan Jews is not surprising, but their clustering so tightly with Ashkenazi rather than Middle Eastern (Iran, Iraq) Jews and not with their host populations is surprising. It is also noteworthy that previously reported mtDNA analysis showed that four maternal lineages underwent dramatic expansion as part of the demographic history of Ashkenazi Jews, and these lineages were also found exclusively among Sephardi Jews, but were not present in non-Jewish samples[18]. These results might suggest either shared ancestry prior to the events leading to the generation of these Diaspora communities, or a previously underappreciated level of contact between these communities. An additional scenario might be Central-Eastern European genetic introgression into Ashkenazi Jews and an Iberian-European gene introgression into Sephardi Jews. According to this formulation, a potentially homogeneous European genetic variation would have generated the genetic proximity between the Jewish communities, with the detailed geographic sources within Europe being different. Finally, it is important to discuss the position of a few individuals outside the cluster. Among Ashkenazi Jews

two samples fall closer to the European populations and comprise two of the three Dutch Ashkenazi Jews included herein. This is not surprising in view of the previously suggested paternal introgression of European males into this Jewish community as evident from the high frequency of Y-chromosome haplogroup R1b among its members[19]. Similarly, a few Moroccan Jewish samples radiate outside the cluster, but are still far from their host population suggesting some gene flow.

The second cluster comprised of the Azerbaijani-Georgian-Iranian-Iraqi-Uzbekistani Jewish samples falls well within the Middle East genetic continuum. The Jewish samples from Iran and Iraq overlap Druze samples but not Iranian samples. The historical events yielding the establishment of the large and long-standing Iraqi and Iranian Jewish communities[16] are addressed extensively by historians; however, the exact demographic mechanisms establishing the smaller Azerbaijani, Georgian and Uzbekistani Jewish communities have received less attention[12]. Our results suggest some genetic proximity between all five Jewish communities comprising this cluster which can be the result of shared genetic origin or continuous interactions during the Diaspora.

The third cluster comprised solely of Yemenite Jews is also evident in the Old World based PC plot and is clearly separated from Yemenites but overlaps with Bedouins and Saudis. This allocation using PCA is interesting as it clearly separates Yemenite Jews from their Yemen hosts, and places them rather with other populations originating or residing in the Arabian Peninsula. The west Eurasian restricted PC plot further supports this observation and suggests a very tight cluster for Yemenite Jews, intermediate in location on the plot between Saudis and Palestinians. Thus, the tight and well- separated Yemenite Jewish samples, suggest a Levantine origin of this community, an early split from other Jewish communities, with strict isolation from

non-Jewish Yemenites. The proximity of the Yemenite Jews to Palestinians and Saudis on the PC plot might reflect a genetic contribution from the Arabian Peninsula, rather than recent shared ancestry. This might not be surprising in view of the historical records suggesting the conversion of Himyarite nobles and others to Judaism and the establishment of a Jewish kingdom in Yemen during the 6th century CE[20,21].

It is noteworthy that the separate designations of these Diaspora communities to the respective clusters discussed herein is rapidly being lost due to the disappearing barrier to admixture in Israel today, such that genotyping of new generations of Israelis should yield some integration of the genetic structure found among the Diaspora Jewish communities designated in the current study.

Supplementary Note 3:

Ancestry analyses, a note on methodology

A recently adapted and widely-used approach to assess the genetic differences/similarities between samples in large multilocus genotyping panels is to use *structure-like* analysis[22]. This increasingly diverse set of Bayesian [23-25] or maximum likelihood (ML) methods[1,26,27] share a common principle in which population structure is inferred as differential membership of individuals in specified number (K) of hypothetical ancestral populations (genetic clusters) characterized by ML estimates for allele frequencies at each loci. When a world-wide sample of individuals is analyzed[1], the assumed number of clusters can correspond to a reference number of distinct divisions, such as continents. Thus individuals can be members of one cluster (e.g. continent) or their genotypes may reflect joint membership in many clusters (e.g. admixture from two or more continents). Because such subdivisions are established geographically or historically, genetic clusters and cluster membership are often discussed in terms of ancestral populations and ancestry admixture proportions.

In fact, both ancestral allele frequencies and ancestry proportions are estimated simultaneously as maximum likelihood estimates for observed genotype data given different input values of K. In most utilizations of this approach, it is the researcher who defines the numbers of ancestral populations – *K*. In practice, the datasets are run multiple times with various assignments of *K* to demonstrate population structure at different levels of subdivision. Each value of K often yields multiple solutions (in terms of underlying putative ancestral frequencies and admixture proportions) predicting the same observed genotype data, therefore it is necessary to explore how many runs converge on the same solution and the likelihood of each.

This type of analysis does not use *a priori* assumptions regarding human population history, therefore its results, if not over-interpreted, are particularly well-suited to uncovering

cryptic structure in large datasets of multilocus genotypes [22]. Most importantly, it should be emphasized that this approach does not model phylogeny. Therefore caution needs to be exercised in drawing inferences regarding the *meaning* of genetic difference and similarity. For example, if two or more different components (putative ancestry signals) are present in a sample, it is not possible to differentiate the relative contribution of recent admixture versus shared ancestry to population structure. However, it would be reasonable to infer recent admixture if individuals within a population vary considerably in their cluster assignments, and *vice versa*. Also, it is not possible from the results of *structure-like analysis* alone to distinguish ancient and more recent splits between populations. For example, differentiation driven by recent random genetic drift of a small isolated population leaves a similar signal as the split of continental populations that occurred during the peopling of Eurasia in a time scale that is many times greater. Therefore, the term 'ancestry' as inferred from structure -like analyses used in the current study, refers to genetic relatedness and should be considered in terms of genetic 'similarity' and 'dissimilarity' irrespective of its genesis, and does not reflect phylogenetic history, though might mirror it.

The ML approaches of structure-like analysis are generally faster than Bayesian, so results for more markers in more individuals are generated in reasonable computing time. Of the ML approaches we chose a recently introduced algorithm assembled into the program *Admixture*[27]. This algorithm is faster and, due to a more stringent convergence criterion, is more accurate than alternatives like *Frappe*[1,27].

We ran ADMIXTURE starting from K=2 upwards in 100 repetitions (using a random seed) using both populations sets – the entire Old World sample set and the West Eurasian restricted sample set (**Fig.3** and **Supplementary Fig. 4**). For the Old World set we observed

geography-driven genetic structure, much like that reported in[1], but with the addition of our new samples, the results are more informative for the Middle East, the Caucasus, North and East Africa, and Eastern Europe.

We studied the Log-likelihood scores (LLs) to assess convergence of ADMIXTURE runs at each value of K. For the global dataset we determined that within the fraction (10%) of runs with the highest LLs the maximum difference between the LLs was minute (<1 LL units), which was five orders of magnitude lower than the average maximum difference of LLs over the 100 runs. Thus, we observed convergence at all levels of K (K=2 to K=10). We note that this approach[28] is indicative rather than definitive for assessing convergence. Applying the same to the West Eurasian dataset suggested that convergence was reached at K=2 to K=5; K=7 and K=8. In general the LLs for any given K occurred on one or more plateaus each corresponding to a different solution for the ML estimate. We demonstrate this by the example of K=7 in the global dataset where three LLs plateaus can be identified (the fourth with the lowest LLs contains too few runs to be considered a "plateau"). We plotted ten random runs from each plateau to show the different ML solutions. In all but one (K=6 of the global dataset) of the converged Ks, the plateau of the highest LLs was also the widest, meaning that majority of the runs yielded the same and probably optimal ML solution (www.evolutsioon.ut.ee/MAIT/additional_plots).

For each converged K we plotted one run from the 10% fraction of runs with the highest LLs (**Supplementary Fig. 4**). In addition we plotted all runs in the 10% fractions to show that indeed the same ML solution was reached each time. Due to size these plots are available on our website (www.evolutsioon.ut.ee/MAIT/additional_plots).

In presenting structure-like analyses the choice of K represents a source of discussion and alternate approaches[23,29]. It is not clear whether the quest for a correct K is meaningful for complex datasets[30] especially in phylogenetic applications. On the other hand, choice of K is more important in accounting for population stratification as a confounding factor in genetic association studies. In our case the progressive addition of components with successive values for K can be viewed as capturing hierarchical population structure with increasing resolution. However, this does not mean that all K values necessarily have demographic historical meaning. Firstly, if there is no real structure in the population studied the model will be fitted to noise. This upper limit for K could be discovered by following informal pointers such as asymmetry of assignment. Here a strong signal for real population structure is inferred if some individuals are strongly assigned to one inferred ancestral population or another, and if the proportions assigned to each group are asymmetric[30]. Another upper limit for K is set by the questions posed in the study. If one is interested in relatedness of the studied populations then K values at which the newly appearing components single out one population are not informative[28]. Note that such structure would still be informative in minimizing population stratification in association studies.

Taking all this into account, we have opted not to restrict the reader with a chosen K but instead present and discuss the full palette of converged Ks (**Supplementary Fig 4**). In the analyses we highlight K=4 of the West Eurasian and K=8 of the global dataset because i) in the West Eurasians dataset components at higher Ks are largely restricted to single populations and ii) in the global dataset most of the major structure is observable already at K=8.

In order to facilitate comparison of inferred ancestry patterns between individuals (and population averages), we performed simple correlation analyses (**Supplementary Fig. 5**). Here we note that whereas positive correlations represent true similarity between the distributions of

the ancestry components of two individuals (or population averages), negative and weak correlations convey an equal signal of difference. Thus, if two individuals, say at K=4, are composed of only two alternative ancestry components and are thus "absolutely" different, they are still 100% similar in not having the remaining two ancestry components.

Supplementary Note 4:

Details of Structure-like Patterns

Old World Analysis (Entire Sample Set)

The Bantu, rather than San/Pygmy specific signal (discernable at K=6) has similar prevalence among the Ethiopian Jews[5], Semitic-speaking Tigreans and Amharas, but is slightly more frequent among the Cushitic-speaking Oromos (**Fig. 3** and **Supplementary Fig. 4**). Presence of sub-Saharan specific component in Moroccans and Egyptians, extending to the near-by populations of the Near East and Arabia, is evident and consistent with a previous report for Algerian Mozabite Berbers, Bedouins, and Palestinians[1]. The main differences between Europe and the Middle East are quantitative, rather than qualitative. At K=4, a correlation matrix reproduces the uniformity of West Eurasia, and the split into the Middle Eastern and European gene pools is evident at K=8 (**Supplementary Fig. 5a**). Note that Cyprus, assigned formally within Europe, coincides with Levantine, not European populations.

At K=8, pairwise "host" – Jewish community correlations shows that the most characteristic difference between Sephardi and Ashkenazi Jews from their historic host populations in Europe, is the limited presence of the light green component among the latter, while dominant in the Middle Eastern and Indus Valley populations (**Supplement Fig. 5a**) . It is noteworthy that this component is virtually absent among the two West European genetic outlier populations - in the French Basque and Sardinians. On the other hand, Moroccan Jews differ from their host population (Moroccan Berbers and Arabs) not only because of the presence of the dominant sub-Saharan African component in the latter, but also because the host populations (e.g., Mozabites) lack the light green component. Moreover, when at K=10 the Moroccans acquire a substantial share of the novel dark violet component, then in the Moroccan Jewish

ancestry component palette it reaches only approximately 10%. These non-Jewish populations, as well as Mozabite Berbers, possess significant levels of the predominantly European dark blue component, which is much less pronounced among the Levantine, Arabian Peninsula, and South Caucasus populations (**Fig. 3**). However, this component is absent in Yemenite Jews – the main characteristic that distinguishes them from all other West Eurasian and North African Jews. Jews from Iran and Iraq appear to be very similar to each other throughout the succession of K values (K2 to K10) (**Supplementary Fig. 4**) and differ from Iranians mostly because the latter have acquired a larger share of the dark green component that starting from K=8 becomes the most dominant component in South Indian populations. Meanwhile, the palette components of individuals of these two Jewish communities are close to those for Druze and other Levantine populations. The profiles of the two Indian Jewish communities, the Bene Israel from Mumbai and even more so the Cochin Jews from the South Indian province of Kerala, are much closer to those of their host populations than to other Jewish communities (**Supplementary Fig. 4**). Nevertheless, at K=8, both of the Indian Jewish populations, in particular the Bene Israel, carry a fraction of the light blue component which is otherwise most prominent in the Middle East (i.e., at more visible levels than in most of the more western Pakistani populations). Furthermore, the Cochini Jews differ markedly from neighboring Dravidic-speaking Southern Indian populations in the presence of the light-green component (**Supplementary Fig. 4**). These western Eurasian components among Indian Jews are consistent with historical records of arrival from western parts of southern Eurasia. Finally, the Uzbekistani Jews differ significantly from Uzbeks, as well as Uygurs and other Central Asians profoundly – insofar as they completely lack the yellow component which is dominant among Han, Japanese, and many others East, Southeast and North Asian populations (**Fig. 3**). This finding is quite remarkable given that Uzbekistani Jews (also

known as Bukharan Jews) have an oral tradition tracing their ancestry back to the deportation of Israelites from the Northern Kingdom by the Assyrians around the mid-eighth Century BCE.

West Eurasian Analysis

We also ran ADMIXTURE on a West Eurasian subset of populations (**Supplementary Table 1**). At K=4, there is a clear separation between the Middle East and Europe (**Supplementary Fig. 4; Supplementary Fig. 5b**). In both regions, genetic structure is dominated by two alternative components with some intermixing along the Mediterranean coast and the Caucasus. Similarly to the PCA results (**Fig. 1** and **Supplementary Fig. 2 and 3**), the Moroccan, Sephardi, and Ashkenazi Jews show a slightly different structure pattern than the other Jewish communities who are virtually indistinguishable from the Levantine populations such as Lebanese, Druze, Syrians, or Samaritans. The former three show closest match to ADMIXTURE component pattern of the Cypriots (**Supplementary Fig. 5b**). The Druze and Samaritans show next best matches for Sephardi and Moroccan Jews while for the Ashkenazi Jews the next best match is the component palette of the Tuscans. The main difference of Moroccan, Sephardi and Ashkenazi Jews from Levantine populations is a higher proportion of the component otherwise predominant in (Mediterranean) West Europe, which comprises as much as approximately 90% of the "palette" of the Sardinians, for example. Simple correlation analyses of the individual component patterns of the Ashkenazi and Sephardi Jews reveals the overall tendency for the former to be slightly more similar to West Europeans than the Sephardi Jews and conversely, for Sephardi Jews to show closer matches to ancestry proportions of the Middle Eastern populations than do the Ashkenazi Jews (Supplementary **Fig. 5c**).

Supplementary Note 5:

Inferences of significance of ASD differences using bootstrap

In **Table 1** of the main text, we show the observed ASD[31] values when comparing Jewish communities to various population groups. We sought to verify which of these differences are indeed statistically significant. Formally, given three populations, for example:

1. A Jewish population: Ashkenazi Jews (denote this population p_1, assume it has n_1 individuals)
2. A host population: Eastern Europeans (p_2, n_2, respectively)
3. A candidate source population: Levant (p_3, n_3)

We sought to test whether:

H_0: $ASD(p_1, p_2) = ASD(p_1, p_3)$ (at the population level)

can be rejected based on the difference between the observed ASD's, which we denote $ASD^{(1)}$, $ASD^{(2)}$. Assume WLOG that $ASD^{(1)} > ASD^{(2)}$

Bootstrap approach

There are two sources of variability in the data: individual sampling and site sampling. A proper bootstrap approach should account for both, i.e., estimate the variance resulting from *simultaneously* sampling individuals and sites. To accommodate this we use a two-dimensional bootstrap approach, where we resample both individuals and sites. Essentially, in each bootstrap iteration we resample n_i individuals from each population i, and resample the sites, creating a new data matrix of the same size as our original matrix, where both rows and columns are resampled, and the original population sizes are preserved. Due to the very high computational

cost of this approach (given the size of our data), we could only practically perform B=100 bootstrap samples.

Our bootstrap approach calculates a standard bootstrap estimate of var($ASD^{(1)} - ASD^{(2)}$) and applies this in a t-test of significance for the observed difference. We feel that this approach is well founded, since ASD values are averages of distances over a very large number of SNPs (while these are not quite independent, there is clearly a lot of independent signal). Hence the central limit theorem should facilitate the use of a simple normal approximation. This is the basis for results we report.

The list of 100 pairs of ASD values we compared and the p-values calculated in both approaches are provided in **Supplementary Table 3**. Cases where the numerator of the ratio above was 0 are marked as $p<0.01$. Because of the very large number of markers, we see that even small differences are often statistically significant.