12-1-2013

No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews

Doron M. Behar
Rambam Health Care Campus, Israel, d_behar@rambam.health.gov.il

Mait Metspalu
Estonian Biocentre, Evolutionary Biology Group, Estonia

Yael Baran
Tel-Aviv University, Israel

Naama M. Kopelman
Tel-Aviv University, Israel

Bayazit Yunusbayev
Estonian Biocentre, Evolutionary Biology Group, Estonia

See next page for additional authors

Recommended Citation

Behar, Doron M.; Metspalu, Mait; Baran, Yael; Kopelman, Naama M.; Yunusbayev, Bayazit; Gladstein, Ariella; Tzur, Shay; Sahakyan, Havhannes; Bahmanimehr, Ardeshir; Yepiskoposyan, Levon; Tambets, Kristiina; Khusnutdinova, Elza K.; Kusniarevich, Aljona; Balanovsky, Oleg; Balanovsky, Elena; Kovacevic, Lejla; Marjanovic, Damir; Mihailov, Evelin; Kouvatsi, Anastasia; Traintaphyllidis, Costas; King, Roy J.; Semino, Ornella; Torroni, Antonio; Hammer, Michael F.; Metspalu, Ene; Skorecki, Karl; Rosset, Saharon; Halperin, Eran; Villems, Richard; and Rosenberg, Noah A., "No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews" (2013). *Human Biology Open Access Pre-Prints.* Paper 41.
http://digitalcommons.wayne.edu/humbiol_preprints/41

Authors
Doron M. Behar, Mait Metspalu, Yael Baran, Naama M. Kopelman, Bayazit Yunusbayev, Ariella Gladstein, Shay Tzur, Havhannes Sahakyan, Ardeshir Bahmanimehr, Levon Yepiskoposyan, Kristiina Tambets, Elza K. Khusnutdinova, Aljona Kusniarevich, Oleg Balanovsky, Elena Balanovsky, Lejla Kovacevic, Damir Marjanovic, Evelin Mihailov, Anastasia Kouvatsi, Costas Traintaphyllidis, Roy J. King, Ornella Semino, Antonio Torroni, Michael F. Hammer, Ene Metspalu, Karl Skorecki, Saharon Rosset, Eran Halperin, Richard Villems, and Noah A. Rosenberg

No Evidence from Genome-Wide Data of a Khazar Origin for the Ashkenazi Jews

Manuscript for *Human Biology*, August 30, 2013

Doron M Behar[1,2,*], Mait Metspalu[2,3,4*], Yael Baran[5], Naama M Kopelman[6], Bayazit Yunusbayev[2,7], Ariella Gladstein[8], Shay Tzur[1], Hovhannes Sahakyan[2,9], Ardeshir Bahmanimehr[9], Levon Yepiskoposyan[9], Kristiina Tambets[2], Elza K. Khusnutdinova[2,10,11], Alena Kushniarevich[2], Oleg Balanovsky[12,13], Elena Balanovsky[12,13], Lejla Kovacevic[14,15], Damir Marjanovic[14,16], Evelin Mihailov[17], Anastasia Kouvatsi[18], Costas Triantaphyllidis[18], Roy J King[19], Ornella Semino[20,21], Antonio Torroni[20], Michael F Hammer[8], Ene Metspalu[3], Karl Skorecki[1,22], Saharon Rosset[23], Eran Halperin[5,24,25], Richard Villems[2,3,26], Noah A Rosenberg[27]

[1]Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa 31096, Israel

[2]Estonian Biocentre, Evolutionary Biology group, Tartu 51010, Estonia

[3]Department of Evolutionary Biology, University of Tartu, Tartu 51010, Estonia

[4]Department of Integrative Biology, University of California, Berkeley 94720, USA

[5]The Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel

[6]Porter School of Environmental Studies, Department of Zoology, Tel-Aviv University, Tel-Aviv 69978, Israel

[7]Institute of Biochemistry and Genetics, Ufa Research Center, Russian Academy of Sciences, Ufa 450054, Russia

[8]ARL Division of Biotechnology, University of Arizona, Tucson, Arizona 85721, USA

[9]Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, Yerevan 0014, Armenia

[10]Institute of Biochemistry and Genetics, Ufa Research Center, Russian Academy of Sciences, Ufa 450054, Russia

[11]Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa 450074, Russia

[12]Vavilov Institute for General Genetics, Russian Academy of Sciences, Moscow 190000, Russia

[13]Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow 115478, Russia

[14]Institute for Genetic Engineering and Biotechnology, Sarajevo 71000, Bosnia and Herzegovina

[15]Faculty of Pharmacy, University of Sarajevo, Sarajevo 71000, Bosnia and Herzegovina

[16]Genos doo, Zagreb 10000, Croatia

[17]Estonian Genome Center, University of Tartu, Tartu 51010, Estonia

[18]Department of Genetics, Development and Molecular Biology, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

[19]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305, USA

[20]Dipartimento di Biologia e Biotecnologie "Lazzaro Spallanzani", Università di Pavia, Pavia 27100, Italy

[21]Centro Interdipartimentale "Studi di Genere", Università di Pavia, Pavia 27100, Italy

[22]Ruth and Bruce Rappaport Faculty of Medicine and Research Institute, Technion-Israel Institute of Technology, Haifa 31096, Israel

[23]Department of Statistics and Operations Research, School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel

[24]Department of Molecular Microbiology and Biotechnology, George Wise Faculty of Life Science, Tel-Aviv University, Tel-Aviv 69978, Israel

[25]International Computer Science Institute, Berkeley, California 94704, USA

[26]Estonian Academy of Sciences, Tallinn 10130, Estonia

[27]Department of Biology, Stanford University, Stanford, California 94305, USA

*These authors contributed equally to this work.

Address for correspondence:

Doron M. Behar	Noah A. Rosenberg
Molecular Medicine Laboratory	Department of Biology
Rambam Medical Center	Stanford University
Haifa, Israel	Stanford, CA, USA
d_behar@rambam.health.gov.il	noahr@stanford.edu

Key words: ancestry, Jewish genetics, population structure, single-nucleotide polymorphisms

Running title: Genetics of Ashkenazi Jewish origins

Abstract. The origin and history of the Ashkenazi Jewish population have long been of great interest, and advances in high-throughput genetic analysis have recently provided a new approach for investigating these topics. We and others have argued on the basis of genome-wide data that the Ashkenazi Jewish population derives its ancestry from a combination of sources tracing to both Europe and the Middle East. It has been claimed, however, through a reanalysis of some of our data, that a large part of the ancestry of the Ashkenazi population originates with the Khazars, a Turkic-speaking group that lived to the north of the Caucasus region ~1,000 years ago. Because the Khazar population has left no obvious modern descendants that could enable a clear test for a contribution to Ashkenazi Jewish ancestry, the Khazar hypothesis has been difficult to examine using genetics. Furthermore, because only limited genetic data have been available from the Caucasus region, and because these data have been concentrated in populations that are genetically close to populations from the Middle East, the attribution of any signal of Ashkenazi-Caucasus genetic similarity to Khazar ancestry rather than shared ancestral Middle Eastern ancestry has been problematic. Here, through integration of genotypes on newly collected samples with data from several of our past studies, we have assembled the largest data set available to date for assessment of Ashkenazi Jewish genetic origins. This data set contains genome-wide single-nucleotide polymorphisms in 1,774 samples from 106 Jewish and non-Jewish populations that span the possible regions of potential Ashkenazi ancestry: Europe, the Middle East, and the region historically associated with the Khazar Khaganate. The data set includes 261 samples from 15 populations from the Caucasus region and the region directly to its north, samples that have not previously been included alongside Ashkenazi Jewish samples in genomic studies. Employing a variety of standard techniques for the analysis of population-genetic structure, we find that Ashkenazi Jews share the greatest genetic ancestry with other

Jewish populations, and among non-Jewish populations, with groups from Europe and the Middle East. No particular similarity of Ashkenazi Jews with populations from the Caucasus is evident, particularly with the populations that most closely represent the Khazar region. Thus, analysis of Ashkenazi Jews together with a large sample from the region of the Khazar Khaganate corroborates the earlier results that Ashkenazi Jews derive their ancestry primarily from populations of the Middle East and Europe, that they possess considerable shared ancestry with other Jewish populations, and that there is no indication of a significant genetic contribution either from within or from north of the Caucasus region.

The Ashkenazi Jewish population has long been a subject of intense scholarly interest from the standpoint of such fields as anthropology, demography, history, medicine, and more recently, genetics. As a result of the availability of high-throughput genetic data covering the whole of the human genome, the last several years have seen major advances in the potential of population genetics to contribute to the study of population relationships and genetic origins (Cavalli-Sforza and Feldman, 2003; Lawson and Falush, 2012; Novembre and Ramachandran, 2011). For the Ashkenazi Jewish population, genetic studies by several different investigators making use of a variety of genetic markers, genotyping platforms, analytical tools, and independently collected samples, have converged on a series of remarkably similar results. First, it is possible to assess whether an individual has Ashkenazi Jewish ancestry, not only for subjects who identify as having exclusively Ashkenazi Jewish ancestors in recent generations, but also, in many cases, for subjects who report only one or two Ashkenazi Jewish grandparents (Bauchet and others, 2007; Guha and others, 2012; Need and others, 2009; Price and others, 2008; Seldin and others, 2006; Tian and others, 2008). Second, Ashkenazi Jewish individuals have relatively long stretches of the genome shared with each other, both in comparison with their genomic sharing with individuals from other populations, and in comparison with levels of within-population genomic sharing in these other populations (Atzmon and others, 2010; Campbell and others, 2012; Guha and others, 2012; Henn and others, 2012). Third, relatively little observable genetic difference exists between representatives of eastern and western Ashkenazi Jewish populations, suggesting that genetically, the Ashkenazi Jewish population approximates a single large community (Guha and others, 2012). Fourth, considering the Ashkenazi Jewish population in relation to other populations, Ashkenazi Jews show the greatest genetic similarity to Sephardi Jews, and, to a lesser extent, to North African Jews (Atzmon and others, 2010; Behar and others, 2010; Campbell and others, 2012; Kopelman and others, 2009).

The issue of the geographic origin of the Ashkenazi Jews has been a source of considerable discussion, repeatedly addressed in the historical literature for over a century (Efron, 2013), and it has similarly not escaped the attention of population genetics. Competing theories include a hypothesis that Ashkenazi Jews descend largely from the Khazar Khaganate, a conglomerate of mostly Turkic tribes, who ruled in what is now southern Russia with the capital Atil in the Volga delta on the northwestern banks of the Caspian Sea approximately 1,400 to 1,000 years ago (**Figure 1**). According to this hypothesis, a portion of the Khazar population, among whom at least some had converted to Judaism, migrated north and west into Europe from their ancestral lands to become the ancestors of some or all of the Ashkenazi Jewish population. This hypothesis can be viewed as an alternative view to a perspective that the Ashkenazi Jewish population originated in the west rather than the east, with Jewish migrations north into Europe from Italy through France. Historical scholarship has provided considerable documentary evidence that Jews did indeed live along this latter route during the period of their entry into central Europe (Baron, 1957; Ben-Sasson, 1976; De Lange, 1984; Mahler, 1971), and the discussion can be viewed as an attempt to evaluate the relative magnitudes of possible eastern and western contributions.

The genetic perspective on Ashkenazi Jewish origins has pointed to a complex and multilayered construction of the Ashkenazi community giving rise to its contemporary shape. Most major genome-wide population-genetic studies of Ashkenazi Jews have detected evidence that the population has elements of ancestry both from Europe and from the Middle East (Atzmon and others, 2010; Behar and others, 2010; Campbell and others, 2012; Kopelman and others, 2009). Ashkenazi Jews have been placed intermediately between non-Jewish Europeans and non-Jewish Middle Easterners in a variety of analyses, including multidimensional scaling and principal components analyses, Bayesian clustering, and population trees. In one of the largest of these studies, encompassing 1,287 subjects from 14 Jewish and 69 non-Jewish populations, we found clear signatures of a Levantine ancestry component for Ashkenazi Jews, a component that was partially shared with other Jewish populations (Behar and others, 2010). These genome-wide results have supported earlier mitochondrial DNA and Y-chromosomal studies, which found that most lineages in the Ashkenazi Jewish population along the male and female lines trace primarily to the Levant, with the remaining lineages likely representing European contributions (Behar and others, 2004; Behar and others, 2006; Behar and others, 2003; Hammer and others, 2009; Hammer and others, 2000; Nebel and others, 2001; Ritte and others, 1993; Santachiara Benerecetti and others, 1993).

Aware of uncertainties in the historical scholarship, genomic studies have also attempted to address the potential Khazar contribution to the Ashkenazi Jewish population, facing the fundamental problem that no contemporary population is identified, either by self-identification or by historians, as Khazars or Khazar descendants. For example, Behar et al. (Behar and others, 2003) suggested that a specific R1a1 Y-chromosomal lineage, comprising 50% of the Ashkenazi Levites and observable in non-Jewish eastern Europeans, could represent either a European contribution or a trace of the lost Khazars. Similarly, based on autosomal markers, Kopelman et al. (2009) (Kopelman and others, 2009), Need et al. (2009) (Need and others, 2009), and Guha et al. (2012) (Guha and others, 2012) detected a small but measurable signal of similarity between Ashkenazi Jews and a sample of the Adygei population from the North Caucasus region. In each of these studies, the possible signal of Caucasus ancestry was relatively small compared to that observed from Europe and the Middle East. However, although no gross signal of Caucasus ancestry has been apparent, it is noteworthy that all of the major genetic studies were able to base their conclusions only on a limited representation of the Caucasus region, thereby leaving open the possibility that such a signal might be detectable in a larger Caucasus sample.

One recent study (Elhaik, 2013), making use of part of our data set (Behar and others, 2010), focused specifically on the Khazar hypothesis, arguing that it has strong genetic support. This claim was built on a series of analyses similar to those performed in our original study that initially reported the data. However, the reanalysis relied on the provocative assumption that the Armenians and Georgians of the South Caucasus region could serve as appropriate proxies for Khazar descendants (Elhaik, 2013). This assumption is problematic for a number of reasons. First, because of the great variety of populations in the Caucasus region and the fact that no specific population in the region is known to represent Khazar descendants, evidence for ancestry among Caucasus populations need not reflect Khazar ancestry. Second, even if it were allowed that Caucasus affinities could represent Khazar ancestry, the use of the Armenians and Georgians as Khazar proxies is particularly poor, as they represent the southern part of the Caucasus region, while the Khazar Khaganate was centered in the North Caucasus and further to the north. Furthermore, among populations of the Caucasus, Armenians and Georgians are geographically the closest to the Middle East, and are therefore expected *a priori* to show the greatest genetic similarity to Middle Eastern populations. Indeed, a rather high similarity of South Caucasus populations to Middle Eastern groups was observed at the level of the whole genome in a recent study (Yunusbayev and others, 2012). Thus, any genetic similarity between Ashkenazi Jews and Armenians and Georgians might merely reflect a common shared Middle Eastern ancestry component, actually providing further support to a Middle Eastern origin of

Ashkenazi Jews, rather than a hint for a Khazar origin.

Here, we examine Ashkenazi Jewish origins by assembling new and previously reported data from the three regions relevant to the origins of the Ashkenazi population, namely, Europe, the Middle East, and the region historically associated with the Khazar Khaganate. The data set, which contains 222 individuals from 13 populations covering the full Caucasus region, as well as 39 individuals from two populations in the region of the Khazar Khaganate located to the north of the Caucasus, is the largest available genome-wide sample set overlapping the Khazar region (**Figure 1**). Our study is the first to integrate genomic data spanning the Khazar region together with a large collection of Jewish samples. With the inclusion of the new data from the region of the Khazar Khaganate, each of a series of approaches, including principal components analysis (PCA), spatial ancestry analysis (SPA), Bayesian clustering analysis, and analyses of genetic distance and identity-by-descent sharing continues to support the view that Ashkenazi Jewish ancestry derives from the Middle East and Europe, and not from the Caucasus region.

Materials and methods

Sample set

All samples reported herein were derived from buccal swabs or blood cells collected with informed consent according to protocols approved by the National Human Subjects Review Committee in Israel and Institutional Review Boards of participating research centers. Individual population assignments follow self-identifications as members of one of the Jewish or non-Jewish populations, at the level of all four grandparents (**Supplemental File 1**).

A total of 1,774 samples, including 352 that are newly reported, were assembled, incorporating 88 non-Jewish populations from Arabia, Central Asia, East Asia, Europe, the Middle East, North Africa, Siberia, South Asia, and Sub-Saharan Africa. The sample collection contains 222 samples representing 13 populations specifically from the Caucasus region and 39 samples representing two populations from the Volga region north to north Caucasus (**Supplemental Table 1**) (Behar and others, 2010; International HapMap and others, 2010; Li and others, 2008; Yunusbayev and others, 2012). A total of 202 samples from 18 Jewish populations spanning the range of the Jewish Diaspora were considered, including 84 novel samples and 118 samples that were previously reported (Behar and others, 2010). The aim of using such a broad data set was to enable analyses of the Ashkenazi Jewish samples to be interpreted in the context of worldwide populations and to specifically allow contrasts of Ashkenazi Jews with populations from three geographic sources that have potentially contributed to their ancestry: Europe, the Middle East, and the geographic regions considered to have been part of the Khazar Khaganate.

It is important to note the conceptual difference between sampling contemporary European, Middle Eastern, and Jewish populations as representing descendants of past populations and suggesting that certain samples might represent the ancient Khazar Khaganate, which disappeared ~1,000 years ago with no apparent modern population representing documented direct Khazar descendants. As it is not possible to rely on known direct descendants of the Khazars, we can merely regard populations presently residing in regions considered to comprise the Khazar Khaganate as potential proxies for Khazar ancestry. Under this assumption, we have employed populations in three geographic regions as possible proxies: South Caucasus (Abkhasian, Armenian, Azeri and Georgian), North Caucasus (Adygei, Balkar, Chechen, Kabardin, Kumyk, Lezgin, Nogai, North Ossetian, and Tabasaran), and the Volga region north of the North Caucasus region (Chuvash and Tatar). Among these three regions, the one considered to best overlap with the center of the Khazar Khaganate is the Volga region, followed by the North Caucasus region. **Supplemental File 1** lists all included regions and populations, the color and three letter codes representing each population throughout the various analyses, and the publication in which they were first used. In addition, when possible, the geographic coordinates assigned for each of the non-Jewish populations are reported.

Genotyping of the new samples

Following the manufacturer's protocol, samples were molecularly analyzed using the Illumina iScan System and the Illumina HumanOmniExpress BeadChip process. Genotype data were evaluated using Illumina GenomeStudio v2011.1, making use of genome build GRCh37/hg19.

Quality control and assembly of the data set

The previously reported data were obtained using five overlapping Illumina genotyping arrays (Human610-Quad, HumanHap650Y, Human660W-Quad, HumanOmniExpress-12v1 730K, and HumanOmni1-Quad), following the manufacturer's protocols, and they were evaluated using GenomeStudio v2011.1 with the latest available manifest files. The raw data from the previously published and new samples were combined first by array version and next lifted using the Liftover tool at the UCSC Genome Browser (Kent and others, 2002) to reflect physical positions of human genome build 37 (GRCh37). Marker rs numbers were matched with dbSNP hg19 build 135 using SNAP (Johnson and others, 2008), and the strand was set according to the 1000 Genomes Project. AT and GC markers were removed in order to minimize potential strand errors during the merging of the data from the different Illumina arrays.

After we merged data from different arrays, the combined data set was filtered using PLINK (Purcell and others, 2007) to include only (i) single-nucleotide polymorphisms (SNPs) with genotyping success rate >99.5% and minor allele frequency >1%, and (ii) individuals with genotyping success rate >96.5%. The stringent genotyping success filter ensures that missing data do not reflect markers that were absent in some of the arrays used less frequently in our panel. After filtering, the data contained 270,898 autosomal SNPs in 1,774 individuals.

We tested for cryptic relatedness in our data set using KING (Manichaikul and others, 2010), finding one cryptic pair of first-degree relatives (both Kurdish Jews), and eight pairs of second-degree cryptic relatives (**Supplemental File 1**). Given the known strong founder effect in some Jewish groups, these pairs were not removed in some of the analyses.

Population groups

Regional population groupings were used for analyses of genetic distance and identity by descent. Where appropriate, some populations were placed into multiple groupings.

1. Middle Eastern Jewish: Azerbaijani Jewish, Georgian Jewish, Iranian Jewish, Iraqi Jewish, Kurdish Jewish, Uzbekistani (Bukharan) Jewish;
2. Sephardi Jewish: Bulgarian Jewish, Turkish Jewish;
3. North African Jewish: Algerian Jewish, Libyan Jewish, Moroccan Jewish, Tunisian Jewish;

4. Middle Eastern: Bedouin, Cypriot, Druze, Jordanian, Lebanese, Palestinian, Samaritan, Syrian;
5. Eastern European: Belorussian, Estonian, Lithuanian, Polish, Romanian, Ukrainian;
6. Western and Southern European: French, Italian, Spanish;
7. North Caucasus: Adygei, Balkar, Chechen, Kabardin, Kumyk, Lezgin, North Ossetian, Tabasaran;
8. South Caucasus: Abkhasian, Armenian, Azeri, Georgian;
9. Caucasus: the union of groups 7 and 8;
10. West Turkic: Azeri, Balkar, Chuvash, Kumyk, Nogai, Tatar;
11. East Turkic: Altaian, Turkmen, Tuvinian, Uygur, Uzbek.

Jewish populations and population groups include "Jewish" in the name, and when "Jewish" is not part of a population or group designation, the population or group is non-Jewish.

A marker subset pruned by linkage disequilibrium patterns

For certain analyses, we thinned the data set to minimize the possible effects of linkage disequilibrium (LD). We used PLINK (Purcell and others, 2007) to calculate an LD score (r^2) for each pair of SNPs in 200-SNP windows, excluding one SNP from the pair if r^2>0.4. The window was advanced by 25 SNPs at a time. This procedure yielded a reduced set of 171,126 SNPs.

Phasing

BEAGLE 3.3.2 (Browning and Browning, 2007) with default parameters was used to phase and impute missing genotypes in the full set of 1,774 samples and 270,898 SNPs. The genotyping error rate was low, 6.5×10^{-4}, with a maximum of 0.032 across individuals, so that relatively few positions were imputed. Positions 20,000,000-40,000,000 of chromosome 6, encompassing the anomalous HLA region, were discarded from the phased data. The phased data were used for both SPA and analyses of identity by descent.

Principal components analysis

SMARTPCA (Patterson and others, 2006) was used to run PCA on the LD-pruned individual data set, and the first three principal components were extracted (**Figure 2a, Supplemental Figures 1 and 2**). No standardization or transformation of genotypes was performed before running SMARTPCA. To present the results at the population level, we show the population median for PC coordinates. PCA results were plotted using R (Team, 2012).

Spatial ancestry analysis

The LOCO-LD localization method (Baran and others, 2013) was used with the phased unpruned data to geographically localize the Jewish samples among the west Eurasian samples (**Figure 2b**). LOCO-LD is an extension of SPA, a recently developed model-based approach for the inference of spatial genetic diversity (Yang and others, 2012). The major improvement that LOCO-LD introduces is a correction for LD between proximal markers. LOCO-LD infers a spatial genetic model by utilizing training samples for which both genotypes and estimated geographic locations are given, and it then uses this model to localize additional samples.

With the current data set, we trained the LOCO-LD model on the non-Jewish samples, and then used the model to localize the Jewish samples. Specifically, the model was trained on samples from western Eurasian populations whose locations are known (**Supplemental File 1**). From each training population, half of the samples were used for training. The inferred parameters of the model were then used to localize the rest of the west Eurasian sample. Thus, the samples localized by LOCO-LD include the other half of the samples from populations of known locations, and samples from populations whose locations are treated as unknown, among them the Jewish samples. We plotted the results using R (Team, 2012), and for clarity we also show median coordinates at the population level.

ADMIXTURE

For analyses with ADMIXTURE (Alexander and others, 2009), a STRUCTURE-like program that distributes individuals across a set of *K* groups inferred from unsupervised mixture-based clustering of multilocus genotypes, we used the LD-pruned unphased data. We ran ADMIXTURE at *K*=2 to *K*=20 clusters, considering 100 replicates for each *K* (**Supplemental Figure 3**).

ADMIXTURE includes a cross-validation procedure to help choose the "best" *K*, defined as the *K* for which the model has the best predictive accuracy (**Supplemental Figure 4**). The approach masks subsets of genotypes and uses the estimated ancestry proportions and allele frequencies under the model to predict the masked genotypes. On the basis of the cross-validation error distribution, the genetic structure in our sample set is best described at *K*=10 (**Figure 3**). To assess the convergence of individual ADMIXTURE runs at each *K*, we monitored the maximum difference in log likelihood (LL) scores in fractions of runs with the highest LL scores at that value of *K*. We assume that a global LL maximum was reached at a given *K* if, say, the 10% of the runs with the highest LL score had minimal ($\lesssim$1 LL unit) variation in LL scores. According to this reasoning, the global LL maximum was reached in runs at *K*=2 to *K*=17, excluding *K*=6, 12, 13, and 16 (**Supplemental Figure 5**). We verified our LL-differences approach using CLUMPP (Jakobsson and Rosenberg, 2007), confirming that indeed all the runs whose LL scores differed by less than 3 from the highest LL score resulted in nearly identical membership proportions (CLUMPP score ≥0.9999) (**Supplemental Figures 6 and 7**).

Judging from the cross-validation error distribution and our assessment of *K* values in which a global maximum likelihood solution was likely reached, we chose *K*=10 as the best single representation of the ADMIXTURE genetic structure of the sample. For convenience, we plotted the runs with the highest LL score (**Figure 3, Supplemental Figure 3**); a nearly identical plot would have resulted had we used any of the runs yielding LL scores within 3 of the best run (as verified by CLUMPP). To facilitate visual inspection of the ADMIXTURE plot at *K*=10, we correlated population-specific average cluster memberships treated as arrays, and plotted, for each Jewish group, the 20 most similar populations (**Figure 4**).

Analysis of allele sharing distance

We calculated allele-sharing distance (ASD) (Gao and Martin, 2009) using the unphased unpruned SNP set (**Figure 5**). We calculated ASD between Ashkenazi Jews and our 11 regional groups. Three separate analyses using different Ashkenazi Jewish groupings were considered: all Ashkenazi Jews (**Figure 5a**), western Ashkenazi Jews only (**Supplemental Figure 8a**), and eastern Ashkenazi Jews only (**Supplemental Figure 8b**). For each computation, we calculated the mean ASD between pairs of individuals, one Ashkenazi Jewish individual and one from the regional group, considering all possible pairs.

To determine whether differences in ASD were statistically significant, we adopted a two-dimensional bootstrap approach (Behar and others, 2010) (**Supplemental Table 1)**. Briefly, we tested a null hypothesis that a difference between two mean ASD values is not significant, by estimating the variance of this difference using a bootstrap approach, and performing a standard normal test with the estimated variance (Behar and others, 2010).

To compare ASD patterns observed with Ashkenazi Jews to those seen with other populations, we repeated the full ASD analysis three times, replacing Ashkenazi Jews with Cypriots, Druze, and Palestinians. For these analyses, Cypriots, Druze, and Palestinians were excluded from their respective regional groups.

Identity-by-descent (IBD) sharing

IBD was analyzed using GERMLINE 1.5.1 (Gusev and others, 2009) on the phased unpruned data. We ran GERMLINE with default parameters (-min_m 3 –bits 128 –err_hom 4 –err_het 1) to detect pairwise IBD sharing for all pairs of study samples. Following previous work (Gusev and others, 2012), we searched for genomic regions in which sparse SNP coverage yields false positive IBD calls, and excised them from the GERMLINE-estimated IBD segments; specially, we divided the genome into non-overlapping 1-Mb blocks and excised blocks with <100 SNPs. We then kept only the shared IBD segments whose length, following excisions, exceeded 3 Mb. Finally, we discarded from the analysis chromosomes 6, 11 and 12, which presented a high level of excessive false-positive sharing, similar to effects observed previously (Gusev and others, 2012).

For each population group *G*, we computed the mean length of IBD sharing with Ashkenazi Jews as $\sum_{i=1}^{N_0}\sum_{j=1}^{N} I_{ij}/(N_0 N)$, where N_0 is the number of Ashkenazi Jewish samples, N is the sample size of group G, and I_{ij} is the total length of shared IBD segments between samples *i* and *j*. To test hypotheses about differential levels of sharing between different groups and the Ashkenazi Jews, we used a Wilcoxon signed rank test. Specifically, let G_1 and G_2 be two population groups. We wish to test the hypothesis that one of these groups shares more IBD segments with the Ashkenazi Jewish group than does the other. For each Ashkenazi Jewish sample *i* we compute $s_{\mathrm{i}1}$ and $s_{\mathrm{i}2}$, the mean IBD sharing between sample *i* and all samples in G_1 and in G_2, respectively. Our null hypothesis is that for every randomly chosen Ashkenazi sample, $P(s_{i1} < s_{i2}) = P(s_{i2} < s_{i1})$. The two-tailed Wilcoxon signed rank *p*-value was computed for each pair (G_1, G_2) (**Supplemental Table 2**).

Results

Principal components analysis

Figure 2a presents the first two principal components (PCs) of genetic variation at three levels of magnification, color-coding the samples by geographic region. The two plots at lower magnification indicate that PC placement of most Jewish populations, including the Ashkenazi Jews, is far from such geographically distant populations as East Asians, South Asians, and Sub-Saharan Africans. In the highest-magnification plot, focusing on the Jewish populations, samples are represented by a three-letter code according to **Supplemental File 1**, and color-coded circles indicate population-level PC coordinate medians. The plot possesses a geographic structure, with Middle Eastern populations at the bottom and European populations at the top, arranged with Southern Europeans on the left and Eastern Europeans on the right.

The Ashkenazi Jewish samples produce a relatively tight cluster that overlaps with some Jewish and non-Jewish populations. Among the Jewish populations, Ashkenazi Jews fall closest to Italian Jews, Middle Eastern Jews, North African Jews, and Sephardi Jews, positioned continuously with other Middle Eastern non-Jewish populations along PC1. Among non-Jewish populations, Ashkenazi Jews lie nearest to Armenians, Cypriots, Druze, Greeks, and Sicilians. Four Ashkenazi Jews fall outside the main Ashkenazi cluster and lie closer to Europeans.

Samples representing the geographic region associated with the Khazar Khaganate are widely spread along the plot. Populations from the northern Volga region (Chuvash and Tatar; see **Figure 1**) are located far from Jewish, Middle Eastern, and Southern European populations and do not appear in the highest-magnification plot focused on the Jewish samples. Populations from the North Caucasus are largely placed in the upper right part of this plot, falling between Turks, Azeris, and Eastern Europeans (**Figure 2a**). The four South Caucasus populations are less closely clustered than the North Caucasus populations, with Azeris overlapping Iranians and Turks, and Abkhasians appearing closer to Eastern Europeans, Kurds, and Turks. The Georgians and Armenians fall close to each other, with Georgians placed between Eastern European populations, North Caucasus populations, Southern Europeans, and the cluster of Ashkenazi Jews, Middle Eastern Jews, Sephardi Jews, Cypriots, and Druze; Armenians lie somewhat closer to this latter cluster, particularly to the tight cluster containing Azerbaijani, Georgian, Iranian, and Kurdish Jews. Within the Khazar region, the farther south a population is, the closer it lies to Middle Eastern non-Jewish populations. No particular similarity of Ashkenazi Jews with Volga or North Caucasus populations is evident; further, the South Caucasus populations fall closer to non-Ashkenazi Middle Eastern Jewish populations than to Ashkenazi Jews.

Spatial ancestry analysis

Figure 2b presents the results of spatial localization with LOCO-LD. Only the samples whose spatial ancestry was inferred with respect to samples of presumed known spatial ancestry are shown. Each sample is placed according to its estimated geographic coordinates, color-coded, and represented by a three letter code according to **Supplemental File 1**. As in the PCA figure, a plot at low magnification indicates placement far from most Jews of populations from a number of distant geographic regions, including Siberians and South Asians.

In the higher-magnification visualization, Ashkenazi Jews form a linear cluster in the latitudinal dimension and are closest to Italian Jews, North African Jews, Sephardi Jews, Cypriots, and Sicilians. Among populations of the Khazar region, as in PCA, the Chuvash and Tatar from the Volga region are absent from the magnified plot, and most North Caucasus and South Caucasus populations appear relatively far from the Jewish populations. Armenians fall closer to the tight cluster of six Middle Eastern Jewish populations, including two from the

Caucasus region (Azerbaijani Jews and Georgian Jews), than to Ashkenazi Jews. Notable again is the lack of any particular similarity between any population group representing one of the three Khazar regions and Ashkenazi Jews. The general pattern seen in PCA is also observed with LOCO-LD, in that South Caucasus populations lie closer than North Caucasus and Volga populations to the Middle East, and that the closest Jewish populations to Armenians are non-Ashkenazi Middle Eastern Jewish groups. Indeed, in relation to other non-Jewish populations, the Armenian median point is nearly equidistant from Middle Eastern and other South Caucasus populations, indicating a general genetic proximity with Middle Eastern populations.

ADMIXTURE

The estimated population structure from ADMIXTURE with K=10 identifies several clusters corresponding to populations that are geographically distant from most Jewish populations, including two clusters centered on Sub-Saharan Africa, two primarily visible in Central Asia and East Asia, and one for the Kalash population from Pakistan (**Figure 3**). Many of the remaining clusters, which we indicate numerically, are spread across broad regions from Europe to South Asia, and it is possible to interpret the placement of Jewish populations in terms of their membership proportions in these clusters.

The Jewish populations separate into five groups with distinct ADMIXTURE patterns (**Figure 3**). First, Indian Jews share similar cluster memberships with other populations of India. A clear link to the Middle East, however, is visible in the presence of clusters $k3$ (light blue) and $k4$ (intermediate blue), both of which appear in the Middle East, and the absence of $k5$ (dark blue), which contributes to Indus Valley populations but is largely missing from the Middle East. Second, Ethiopian Jews are similar to their geographic neighbors, with membership proportions that are largely indistinguishable from Amhara and Tigray Semitic-speaking Ethiopians. Third, Yemenite Jews separate from other Jews, with increased membership in cluster *k*3. Fourth, Caucasus (Azerbaijani and Georgian) and Middle Eastern (Iranian, Iraqi, Kurdish and Uzbekistani) Jews form a group, with similar membership proportions in clusters *k*3, *k*4, and *k*7 (dark green). Finally, the fifth and largest Jewish group unites Ashkenazi, North African, and Sephardi Jews. While these populations do differ slightly in the proportions of clusters *k*2 (light red), *k*4, and *k*5, their genetic similarity is striking. Minimal distinction is visible between the Western and Eastern Ashkenazi Jews, but a minutely elevated membership is visible in the Eastern Ashkenazi group for the largely East Asian clusters *k*9 (yellow) and *k*10 (orange).

To complement the visual assessment of clustering patterns, we next identified populations with patterns most similar to Jewish groups by quantitatively correlating population-specific mean membership proportions, treated as arrays (**Figure 4**). For the Jewish populations included in a large group containing Ashkenazi, North African, and Sephardi Jews, most of the populations with the highest similarity of cluster membership coefficients are other Jewish populations. Considering ten Jewish populations included in the group (Algerian, Belmonte, Bulgarian, Eastern Ashkenazi, Italian, Libyan, Moroccan, Tunisian, Turkish, Western Ashkenazi), the non-Jewish populations that appear on lists of populations with the most similar cluster memberships are French Basques, Bulgarians, Cypriots, Druze, Greeks, Italians from Abruzzo, Bergamo, Sicily, and Tuscany, Jordanians, Lebanese, Palestinians, Samaritans, Italians from Sardinia, Spanish, and Syrians. Notably absent from this list is the inclusion of any of the populations from the Khazar region. Among Jewish populations, the only groups for which any Caucasus non-Jewish populations are among the populations in **Figure 4** with the greatest clustering similarity are the Jewish populations from the Caucasus (Azerbaijani and Georgian) and the Middle East (Iranian, Iraqi, Kurdish, Syrian, and Uzbekistani). For these groups, with the exception of the 19th and 20th most similar populations to the Uzbekistani Jews, within the Khazar region, the list of the closest populations ordered by clustering similarity includes only South Caucasus populations.

Genetic distance analysis

The mean ASD differences from 11 population groups were similar for comparisons examining all Ashkenazi Jews, Western Ashkenazi Jews and Eastern Ashkenazi Jews (**Supplemental Figure 8**, **Supplemental Table 1**), and we thus concentrate here on the analysis of all Ashkenazi Jews. The lowest mean ASD was to Sephardi Jews (0.2721), followed by Western and Southern Europeans, South Caucasus populations, Eastern Europeans, and North African Jews. The mean ASD from the Middle East was 0.2767, among the largest values. **Figure 5a** shows a density plot of the pairwise ASD values between Ashkenazi Jewish individuals and individuals in each group, producing the same patterns as those seen with the mean values.

Figures 5b-d respectively show density plots of pairwise ASD between Cypriots, Druze and Palestinians and individuals in each of the 11 groups (excluding the tested population from the relevant group). These plots identify the South Caucasus among the closest of the 11 groups to Cypriots, Druze, and Palestinians as was observed for Ashkenazi Jews. This pattern reflects the observation seen in other analyses that signals of close genetic proximity to the South Caucasus are observed in Middle Eastern populations and are not unique to Ashkenazi Jews.

Supplemental Table 1 shows differences between mean ASDs and *p*-values for the null hypothesis that no difference exists. Except for the smallest differences, most differences are statistically significant. For example, mean ASD between Ashkenazi and Sephardi Jews is smaller than mean ASD between Ashkenazi Jews and Western and Southern Europeans by a non-significant 0.00043 (*p*=0.18); this ASD is smaller than the mean ASD between Ashkenazi Jews and the South Caucasus by 0.0013, but the larger difference is significant (*p*=0.0018).

Identity-by-descent sharing

Figure 6 reports the mean genomic sharing between Ashkenazi Jews and the 11 population groups, and **Supplemental Table 2** gives *p*-values for tests of the null hypotheses of equal mean IBD sharing with Ashkenazi Jews for pairs of population groups. The greatest level of sharing was observed with Sephardi Jews, considerably greater than with other populations. Substantial sharing with Eastern Europeans was also observed, though at a much lower level. Sharing with most other populations was lower still, and with Caucasus populations, the level of sharing was similar to that observed for the Middle East. In accordance with the results from other analyses, the IBD sharing of Caucasus populations with Ashkenazi Jews was relatively low.

Discussion

This work has been the first to assemble extensive genome-wide data from all three regions that have been proposed as ancestral sources for the Ashkenazi Jewish population (**Figure 1**). The collection of samples from contemporary European, Middle Eastern, and Jewish populations is straightforward, as multiple forms of documentation, including the cultural identities of the populations themselves, link the modern populations to ancestral groups living at the time of the early history of the Ashkenazi Jews. By contrast, obtaining samples representing Khazars, for whom no direct link to extant populations has been established, mandates careful consideration. Recognizing this problem, we proceeded by including as many samples as possible from a region encompassing the geographic range believed to correspond to the Khazar Khaganate. After assembly of the data set, we focused our analysis on the geographic origin of the Ashkenazi Jewish population, employing a variety of analyses of population-genetic structure.

Population-genetic structure and Ashkenazi Jews

Our sample set representing the geographic region of the Khazar Khaganate can be split into three subsets (**Figure 1**): populations from the South Caucasus region (Abkhasian, Armenian, Azeri, Georgian), populations from the North Caucasus region (Adygei, Balkar, Chechen, Kabardin, Kumyk, Lezgin, Nogai, North Ossetian, Tabasaran), and populations from the Volga region in the most northerly reaches of the Khazar expanse (Chuvash, Tatar). Under the hypothesis of a strong Khazar contribution to the Ashkenazi Jewish population, we might have expected in PCA (**Figure 2a**) to see the Ashkenazi Jews placed in tight overlap with populations representing the Khazar region. Instead, considering the samples of the Khazar region together with the Ashkenazi Jewish samples, the Ashkenazi Jews were positioned alongside other Jewish samples, between Southern Europeans and samples from the Middle East, and they did not substantially overlap populations from the Khazar region. The three subsets of the Khazar region —South Caucasus, North Caucasus, and Volga region—are themselves differentiated in PCA, with the Volga and North Caucasus populations, which approximate the Khazar region more closely than do the South Caucasus populations, positioned most distantly from Ashkenazi Jews.

Whereas PCA is an unsupervised approach for placing samples in a low-dimensional space, treating all populations as having unknown coordinates *a priori*, LOCO-LD spatial ancestry analysis represents a supervised approach in which Jewish populations are placed in a spatial diagram in relation to non-Jewish samples whose geographic locations are treated as known (**Figure 2b**). The spatial ancestry analysis confirms and sharpens the lack of evidence for the Khazar hypothesis observed in PCA, placing the Ashkenazi Jewish sample in close proximity to Italian Jews, North African Jews, Sephardi Jews, and Mediterranean non-Jewish populations such as Cypriots and Italians. Of the three sub-regions of the Khazar Khaganate, the two northern groups are again distant from the Ashkenazi Jews. Among the four South Caucasus populations, the Armenian and Azeri populations in particular lie closer to non-Jewish Middle Eastern populations, including Druze, Iranians, Kurds, and Lebanese, than to Ashkenazi Jews. Strikingly, the Ashkenazi Jewish population shows no overlap even with the South Caucasus groups, and moreover, it is apparent that the South Caucasus Armenian population is genetically closer to Middle Eastern Jewish populations than to Ashkenazi Jews.

Our principal components (**Figure 2a**) and spatial ancestry analyses (**Figure 2b**) both highlight two pairs of Ashkenazi Jewish samples distant from the major cluster of Ashkenazi Jews. The first pair lies within the larger European cluster and consists of two Ashkenazi Jewish samples from the Netherlands population, which was previously shown to be admixed at the level of uniparental markers (Behar and others, 2004). The other pair includes a Belorusian and a

Romanian sample that are genetically outside the Jewish and European clusters, likely representing recent undocumented admixture rather than a signal of ancient Khazar origin. Because time erodes the variance of admixture across individuals in admixed populations (Verdu and Rosenberg, 2011), had a Khazar contribution been made to the Ashkenazi community ~1,000 years ago, it would be common to nearly all modern Ashkenazi individuals through generations of endogamy, and would not be centered on a few outliers.

We used a maximum likelihood based STRUCTURE-like approach as implemented in ADMIXTURE to assess the position of the Jewish groups in relation to the established genetic structure of Eurasian populations (Auton and others, 2009; Behar and others, 2010; Li and others, 2008; Metspalu and others, 2011; Yunusbayev and others, 2012) (**Figure 3**). In this analysis, a large group of Jewish populations, containing Ashkenazi, North African, and Sephardi Jews, produced similar patterns of membership. The similarity of the genetic membership proportions suggests a common origin of the Jewish populations in this group and limited or comparable levels of admixture with closely related host populations. Similar membership in the *k*5 component might be interpreted as an admixture event between Jews and European host populations that predates the split of European and North African Jews.

Genetic structure is evident within the larger group of similar Jewish populations. North African Jews show slightly elevated membership in the *k*2 component prevalent in African populations. Similarly, in the Ashkenazi Jews, the proportion of the largely European *k*5 component is somewhat larger than that in the Sephardi Jews (23% vs. 16%). Within the Ashkenazi Jews from Eastern and Central Europe, we do see a signal (2.2%) of components common in East Asia that are less visible in Ashkenazi Jews from Western Europe or European Sephardi Jews (0.6%). These components also appear in Eastern Europeans and in some Middle

Eastern populations, such as Yemenis, so that it is difficult to attribute their minor elevation in Eastern Ashkenazi Jews to a particular origin. The most prevalent cluster in the Caucasus region is cluster *k*6, which appears throughout Europe, the Middle East, and South Asia. Nevertheless, the Ashkenazi Jews do not stand out from other Jewish populations in possessing higher proportions of this component. Rather, Caucasus Jews and even Sephardi Jews have higher proportions of the component dominant in the Caucasus than do Ashkenazi Jews.

In brief, judging from the similarity of the membership proportion distributions (**Figure 4**), ADMIXTURE demonstrates the connection of Ashkenazi, North African, and Sephardi Jews, with the most similar non-Jewish populations to Ashkenazi Jews being Mediterranean Europeans from Italy (Sicily, Abruzzo, Tuscany), Greece and Cyprus. When subtracting the *k*5 component, which perhaps originates in Ashkenazi and Sephardi Jews from admixture with European hosts, the best matches for membership patterns of the Ashkenazi Jews shift to the Levant: Cypriots, Druze, Lebanese, and Samaritans.

Quantitative measures of genetic proximity from genetic distance analysis agree with the results of the other methods (**Figure 5**). As no significant differences in genetic distance (**Supplemental Figure 8**, **Supplemental Table 1**) were noted when the Ashkenazi Jews were split into Eastern and Western groups, in agreement with the work of Guha et al. (2012) (Guha and others, 2012), we examined Ashkenazi Jews as a single population. The lowest mean genetic distance for Ashkenazi Jews was with Sephardi Jews, followed by Western and Southern Europeans, populations of the South Caucasus, and North African Jews. Repeating the analysis by comparing the most relevant non-Jewish Middle Eastern populations to all other groups, we found that the greatest proximity of Middle Eastern populations was to the South Caucasus, again suggesting that any Ashkenazi similarity to the South Caucasus merely reflects a Middle Eastern component of their ancestry (**Supplemental Table 1**).

Analysis of genomic sharing, focused on IBD sharing between Ashkenazi Jews and population groups, further sharpens the results from genetic distance analysis (**Figure 6**). IBD analysis, which is focused on the most recent tens of generations of ancestry, is expected to generate tighter clustering of individuals within populations, between populations that have a recent common ancestral deme, or between populations that have recently experienced reciprocal gene flow (Gusev and others, 2009; Gusev and others, 2012). Considering the IBD threshold of 3 Mb for shared segments, Ashkenazi Jews are expected to show no significant IBD sharing with any population from which they have been isolated for ~20 generations. In accordance with the results from the other methods of analysis, Ashkenazi Jews show significant IBD sharing only with Eastern Europeans, North African Jews, and Sephardi Jews. Sharing was minimal with Middle Eastern populations, a not unexpected result given that the time frame for the split from Middle Eastern populations is beyond the detection power of our IBD analysis.

Conclusions: no evidence for a Khazar origin

Cumulatively, our analyses point strongly to ancestry of Ashkenazi Jews primarily from European and Middle Eastern populations and not from populations in or near the Caucasus region. The combined set of approaches suggests that the observations of Ashkenazi proximity to European and Middle Eastern populations in population structure analyses reflect actual genetic proximity of Ashkenazi Jews to populations with predominantly European and Middle Eastern ancestry components, and lack of visible introgression from the region of the Khazar Khaganate—particularly among the northern Volga and North Caucasus populations—into the Ashkenazi community. We note that while we find no evidence for any significant contribution of the Khazar region to the Ashkenazi Jews, we cannot rule out the possibility that a level of Khazar or other Caucasus admixture occurred below the level of detectability in our study. Contemporary populations represent the outcome of many layers of minor and major demographic events that do not always leave a visible genetic signature. However, our study clearly identifies signals of Europe and the Middle East in Ashkenazi Jewish ancestry, rendering any possible undetected Khazar contribution below a minimal threshold.

Our results contrast sharply with the work of Elhaik (Elhaik, 2013), which claimed strong support for a Khazar origin of Ashkenazi Jews. This disagreement merits close examination. Elhaik (Elhaik, 2013) based his claim for Khazar ancestry of the Ashkenazi Jewish population on an assumption that two South Caucasus populations, Georgians and Armenians, are suitable proxies for Khazar descendants, and on observations of similarity of these populations with Ashkenazi Jews. By assembling a larger data set containing populations that span the full range of the Khazar Khaganate, we find no evidence that a particular similarity exists between Ashkenazi Jews and any of the populations of the Khazar region; further, within the region, the newly incorporated northern populations that best overlap with the presumed center of the Khazar Khaganate are the most genetically distant from Ashkenazi Jews.

While we do observe some evidence of similarity between Ashkenazi Jews and South Caucasus populations, particularly the Armenians, it is important to assess whether this similarity could reflect Khazar origins or might merely reflect a shared ancestry of Ashkenazi Jews and South Caucasus populations in the Middle East. We find that the Ashkenazi Jews carry no particular genetic similarity to the South Caucasus any more than do many other populations from the Middle East, Mediterranean Europe, and particularly, several of the Middle Eastern Jewish populations. The South Caucasus has been previously shown (Haber and others, 2013;

Yunusbayev and others, 2012) and here again to have common genetic ancestry with much of the Middle East. Therefore, it cannot be claimed that evidence of Ashkenazi Jewish similarity to Armenians and Georgians reflects a South Caucasus origin for Ashkenazi Jews without also claiming that the same South Caucasus ancestry underlies both Middle Eastern Jews and a large number of non-Jewish populations both from the Middle East and from Mediterranean Europe. Thus, if one accepts the premise that similarity to Armenians and Georgians represents Khazar ancestry for Ashkenazi Jews, then by extension one must also claim that Middle Eastern Jews and many Mediterranean European and Middle Eastern populations are also Khazar descendants. This claim is clearly not valid, as the differences among the various Jewish and non-Jewish populations of Mediterranean Europe and the Middle East predate the period of the Khazars by thousands of years (Baron, 1957; Ben-Sasson, 1976; De Lange, 1984; Mahler, 1971).

We take this opportunity to clarify the differences between genetic proximity of populations and ancestor-descendant relationships. Most illustrative are the results obtained from ADMIXTURE (**Figure 3**). This analysis clearly shows that throughout the vast western Eurasian region, populations share the same genetic clusters, albeit at different frequencies. These genetic components typically represent ancient genetic forces that have shaped the current genetic landscape, and an attempt to connect them to a particular population that has likely arisen much later than their establishment is inherently problematic. Specifically, our analysis highlighted the Armenian population, and to a lesser extent, the Azeri population, as the only Caucasus populations that present all genetic components also observed in Middle Eastern and North African populations. Thus, Armenians, used by Elhaik (Elhaik, 2013) as a potential proxy for a Khazar source population, could equally well have been employed as a misleading proxy for many populations across the Middle East with similar cluster memberships, thereby producing the same problematic interpretation that each such population is ancestral to Ashkenazi Jews. The mere finding of shared cluster membership does not unambiguously attest to a demographic event responsible for the cluster and, therefore, cannot be further interpreted to suggest that one population is ancestral to another population simply because it is found within the same cluster. Thus, for example, it would be misleading to conclude from the ADMIXTURE analysis that Ashkenazi Jews are actually the primary source population giving rise to the Sicilians, Druze, or North African Jews with whom they share similar membership coefficients.

In summary, in this most comprehensive study to date, we have examined the three potential sources for contemporary Ashkenazi Jews, using a new sample set that covers the full extent of the Khazar realm of the 6th to 10th centuries. Analysis of this large data set does not change and in fact reinforces the conclusions of multiple past studies, including ours and those of other groups (Atzmon and others, 2010; Bauchet and others, 2007; Behar and others, 2010; Campbell and others, 2012; Guha and others, 2012; Haber and others, 2013; Henn and others, 2012; Kopelman and others, 2009; Seldin and others, 2006; Tian and others, 2008). We confirm the notion that the Ashkenazi, North African, and Sephardi Jews share substantial genetic ancestry and that they derive it from Middle Eastern and European populations, with no indication of a detectable Khazar contribution to their genetic origins.

Acknowledgments

We acknowledge the Morrison Institute for Population and Resource Studies for travel support for DMB to visit the lab of NAR. RV, MM, EM, KT, BY, AK, and HS were supported by European Union European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre, and University of Tartu and Estonian Basic Research grant SF0270177As08. MM thanks Estonian Science Foundation grant nr 8973 and Baltic-American Freedom Foundation Research Scholarship program. EH is a faculty fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. YB was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. EH was also partially supported by US National Science Foundation grant III-1217615. We thank the High Performance Computing Center of the University of Tartu and Tuuli Mets from the Estonian Biocentre core facility.

References

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655-1664.

Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E and others. 2010. Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. Am J Hum Genet 86(6):850-859.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN and others. 2009. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res 19(5):795-803.

Baran Y, Quintela I, Carracedo A, Pasaniuc B, Halperin E. 2013. Enhanced Localization of Genetic Samples through Linkage-Disequilibrium Correction. Am J Hum Genet.

Baron SW. 1957. A social and religious history of the Jews: The Jewish Publication Society of America, Philadelphia.

Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. 2007. Measuring European population stratification with microarray genotype data. Am J Hum Genet 80(5):948-956.

Behar DM, Garrigan D, Kaplan ME, Mobasher Z, Rosengarten D, Karafet TM, Quintana-Murci L, Ostrer H, Skorecki K, Hammer MF. 2004. Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum Genet 114(4):354-365.

Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, Tzur S, Pereira L, Amorim A, Quintana-Murci L, Majamaa K and others. 2006. The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. Am J Hum Genet 78(3):487-497.

Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D and others. 2003. Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. Am J Hum Genet 73(4):768-779.

Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G and others. 2010. The genome-wide structure of the Jewish people. Nature 466(7303):238-242.

Ben-Sasson HH. 1976. A history of the Jewish people. Harvard University Press C, MA, editor.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084-1097.

Campbell CL, Palamara PF, Dubrovsky M, Botigue LR, Fellous M, Atzmon G, Oddoux C, Pearlman A, Hao L, Henn BM and others. 2012. North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. Proc Natl Acad Sci U S A 109(34):13865-13870.

Cavalli-Sforza LL, Feldman MW. 2003. The application of molecular genetic approaches to the study of human evolution. Nat Genet 33 Suppl:266-275.

De Lange N. 1984. Atlas of the Jewish World: Phaidon Press.

Efron JM. 2013. Jewish genetic origins in the context of past historical and anthropological inquiries. Hum Biol In press.

Elhaik E. 2013. The missing link of Jewish European ancestry: contrasting the Rhineland and the Khazarian hypotheses. Genome Biol Evol 5(1):61-74.

Gao X, Martin ER. 2009. Using allele sharing distance for detecting human population stratification. Hum Hered 68(3):182-191.

Golden PB, Ben-Shammai H, Róna-Tas A. 2007. The World of the Khazars: New Perspectives. Selected Papers from the Jerusalem 1999 International Khazar Colloquium. Golden PB, Ben-Shammai H, Róna-Tas A, editors. Leiden Boston: Brill.

Guha S, Rosenfeld JA, Malhotra AK, Lee AT, Gregersen PK, Kane JM, Pe'er I, Darvasi A, Lencz T. 2012. Implications for health and disease in the genetic signature of the Ashkenazi Jewish population. Genome Biol 13(1):R2.

Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I. 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res 19(2):318-326.

Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, Pe'er I. 2012. The architecture of long-range haplotypes shared within and across populations. Mol Biol Evol 29(2):473-486.

Haber M, Gauguier D, Youhanna S, Patterson N, Moorjani P, Botigue LR, Platt DE, Matisoo-Smith E, Soria-Hernanz DF, Wells RS and others. 2013. Genome-wide diversity in the levant reveals recent structuring by culture. PLoS Genet 9(2):e1003316.

Hammer MF, Behar DM, Karafet TM, Mendez FL, Hallmark B, Erez T, Zhivotovsky LA, Rosset S, Skorecki K. 2009. Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. Hum Genet 126(5):707-717.

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, Oppenheim A, Jobling MA, Jenkins T and others. 2000. Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. Proc Natl Acad Sci U S A 97(12):6769-6774.

Henn BM, Hon L, Macpherson JM, Eriksson N, Saxonov S, Pe'er I, Mountain JL. 2012. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 7(4):e34267.

International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F and others. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52-58.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23(14):1801-1806.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24(24):2938-2939.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res 12(6):996-1006.

Kopelman NM, Stone L, Wang C, Gefel D, Feldman MW, Hillel J, Rosenberg NA. 2009. Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. BMC Genet 10:80.

Lawson DJ, Falush D. 2012. Population identification using genetic data. Annu Rev Genomics Hum Genet 13:337-361.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL and others. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100-1104.

Mahler RA. 1971. History of Modern Jewry: Schocken.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867-2873.

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Magi R, Metspalu E, Remm M and others. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet 89(6):731-744.

Nebel A, Filon D, Brinkmann B, Majumder PP, Faerman M, Oppenheim A. 2001. The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. Am J Hum Genet 69(5):1095-1112.

Need AC, Kasperaviciute D, Cirulli ET, Goldstein DB. 2009. A genome-wide genetic signature of Jewish ancestry perfectly separates individuals with and without full Jewish ancestry in a large random sample of European Americans. Genome Biol 10(1):R7.

Novembre J, Ramachandran S. 2011. Perspectives on human population structure at the cusp of the sequencing era. Annu Rev Genomics Hum Genet 12:245-274.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet 2(12):e190.

Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, Ruiz-Linares A, Groop L, Saetta AA, Korkolopoulou P and others. 2008. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4(1):e236.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and others. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559-575.

Ritte U, Neufeld E, Prager EM, Gross M, Hakim I, Khatib A, Bonne-Tamir B. 1993. Mitochondrial DNA affinity of several Jewish communities. Hum Biol 65(3):359-385.

Santachiara Benerecetti AS, Semino O, Passarino G, Torroni A, Brdicka R, Fellous M, Modiano G. 1993. The common, Near-Eastern origin of Ashkenazi and Sephardi Jews supported by Y-chromosome similarity. Ann Hum Genet 57(Pt 1):55-64.

Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, Belmont JW, Klareskog L, Gregersen PK. 2006. European population substructure: clustering of northern and southern populations. PLoS Genet 2(9):e143.

Team RC. 2012. R: A language and environment for statistical computing. Vienna, R Foundation for Statistical Computing.

Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi L, Gregersen PK and others. 2008. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4(1):e4.

Verdu P, Rosenberg NA. 2011. A general mechanistic model for admixture histories of hybrid populations. Genetics 189(4):1413-1426.

Yang WY, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. Nat Genet 44(6):725-731.

Yunusbayev B, Metspalu M, Jarve M, Kutuev I, Rootsi S, Metspalu E, Behar DM, Varendi K, Sahakyan H, Khusainova R and others. 2012. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol 29(1):359-365.

Figure Legends

Figure 1: Geographical map of the populations included in this study. The crude borders of the Khazar Khaganate at three stages of its expansion, along with its capital at Atil, are shown. The Khazar Khaganate (~650-1000 CE), one of the largest states of medieval Eurasia, extended from the Volga region in the north to the Northern Caucasus and Crimea in the south, and from present-day Ukraine in the west to the western borders of present-day Kazakhstan and Uzbekistan in the east (Golden and others, 2007). Abbreviations as well as exact geographic locations are detailed in **Supplemental File 1**.

Figure 2: Principal components analysis and spatial ancestry analysis. a. The first two principal components, shown at three stages of magnification of the same plot of all individuals included in this study. Each letter code (**Supplemental File 1**) corresponds to one individual, and the color indicates the geographic region of origin. Median coordinate points for populations are shown as circles. **b**. Scatter plot results of the inferred locations of all individuals in relation to the actual geographic locations of the reported populations. As in part a, each letter code (**Supplemental File 1**) corresponds to one individual, the color indicates the geographic region of origin, and median coordinate points for populations are shown as circles.

Figure 3: Population structure inferred by ADMIXTURE analysis. ADMIXTURE plots at *K*=10 (see methods for choice of *K*). Each individual is represented by a vertical (100%) stacked column of genetic membership proportions. The Jewish groups are highlighted in red. Western Ashkenazi Jews: France, Germany, Netherlands; Central and Eastern

Ashkenazi Jews: Austria, Belorussia, Hungary, Latvia, Lithuania, Poland, Romania, Russia. See **Supplemental Figure 3** for plots at other values of *K*.

*Includes Altaians from Southern Siberia;

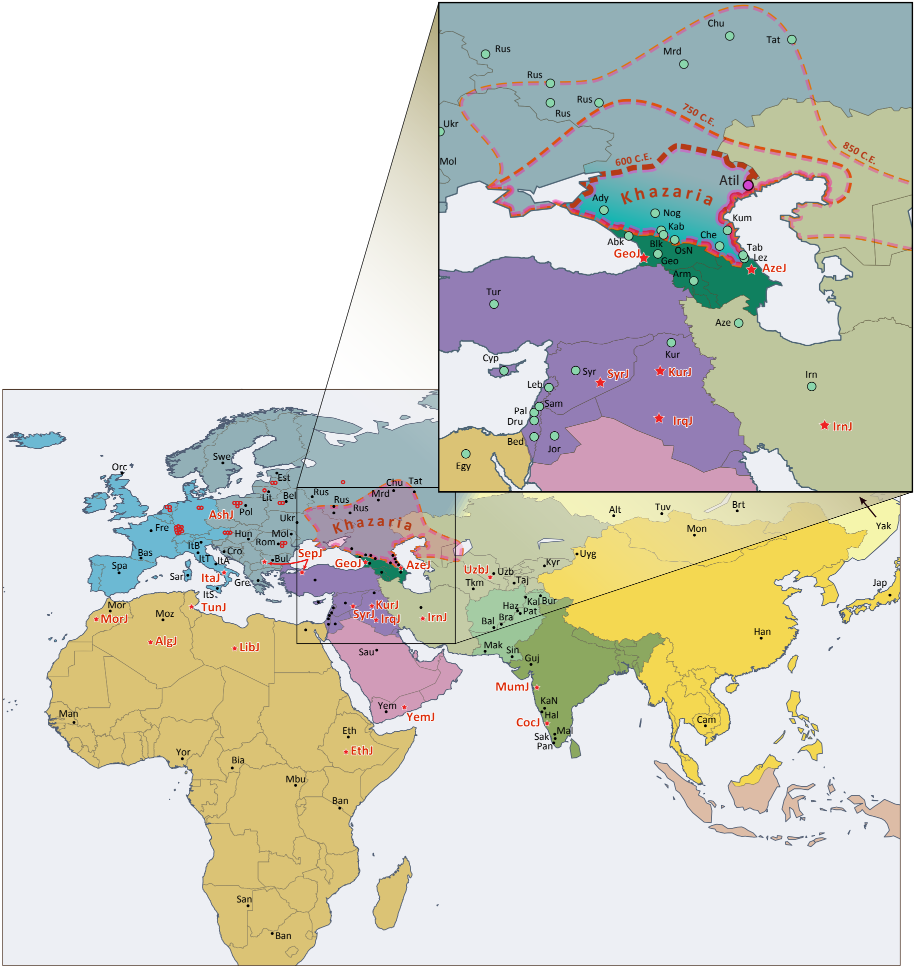
§Belmonte Jewish;

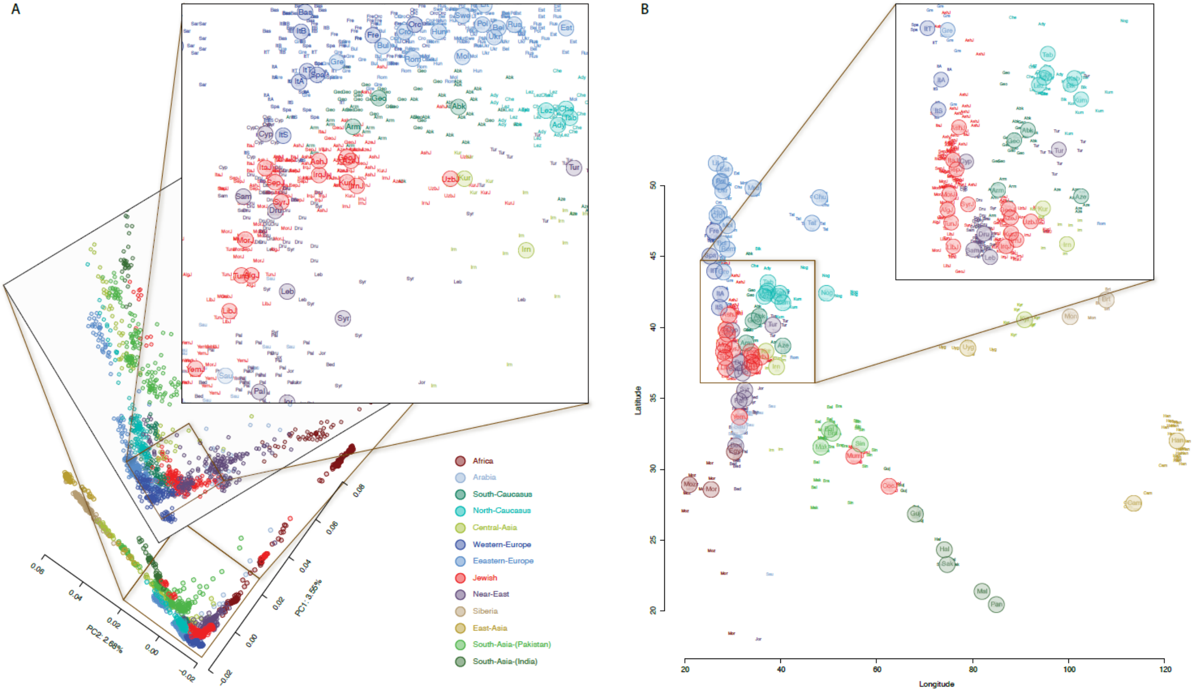
¶Syrian Jewish.

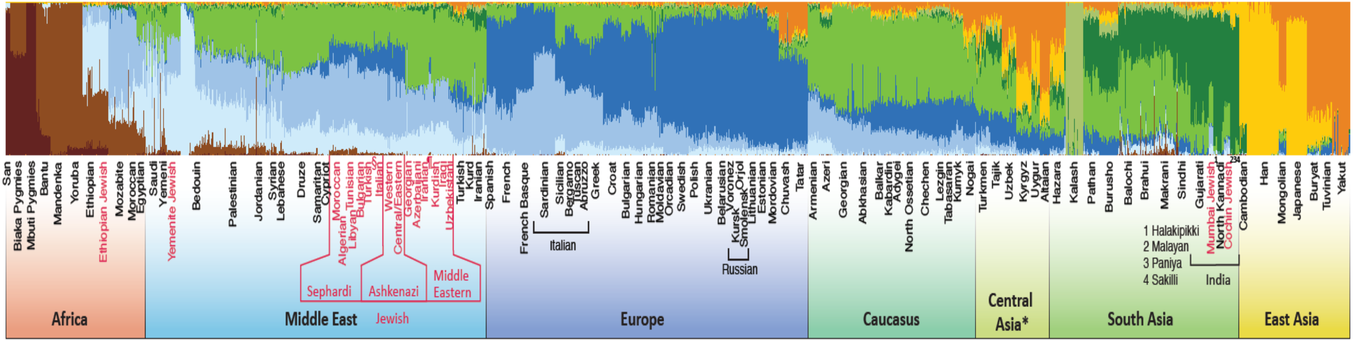
Figure 4: Correlation of population-level mean membership proportions. Each plot, based on the proportions inferred in Figure 3, shows the populations with the highest correlation of membership proportions (up to 20 populations, if the number of populations with a high value of the correlation is large).

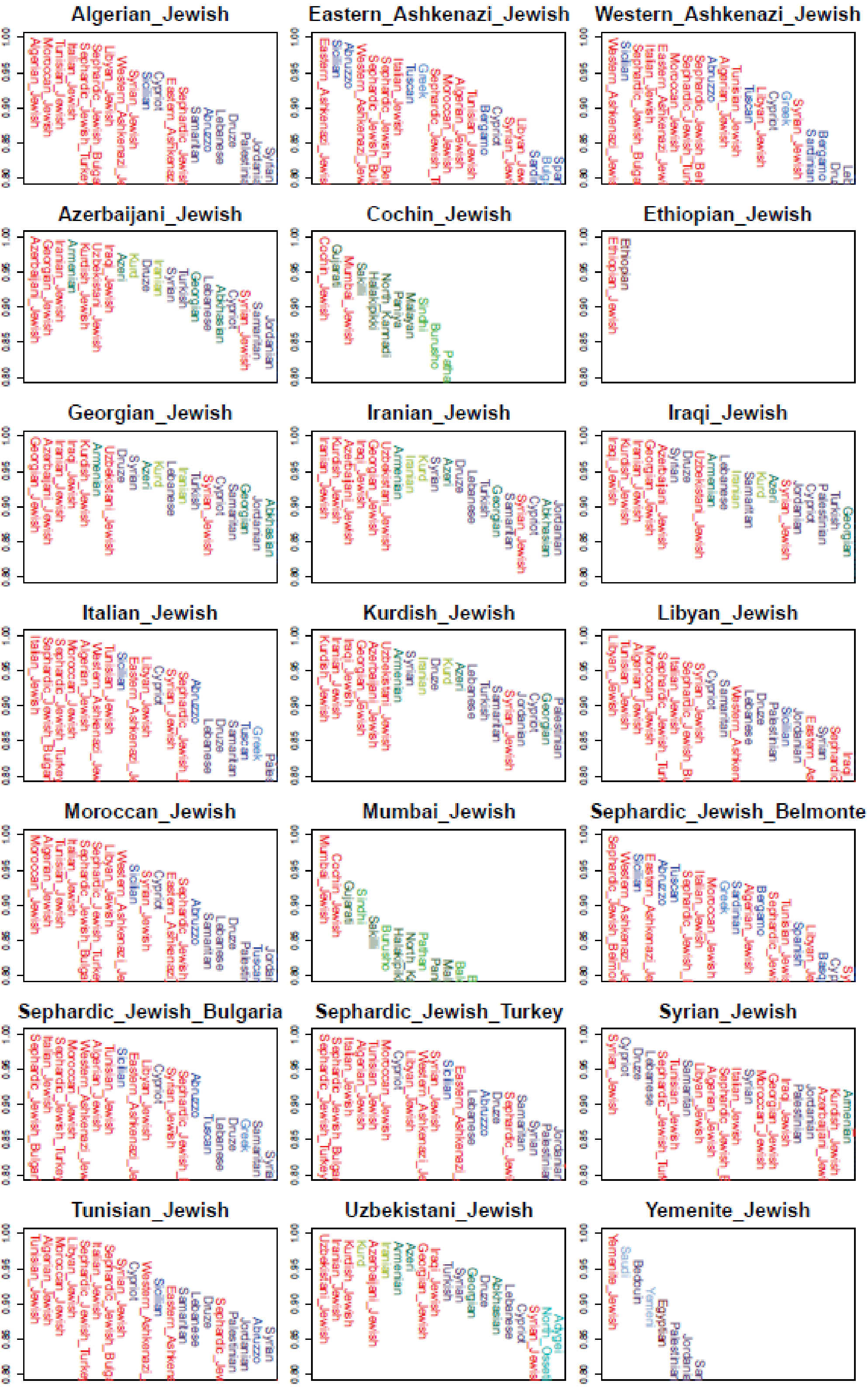
Figure 5: Density plot of pairwise genetic distances between individuals from a specific population and individuals from population groups. a. Ashkenazi Jews. b. Cypriots. c. Druze. d. Palestinians.

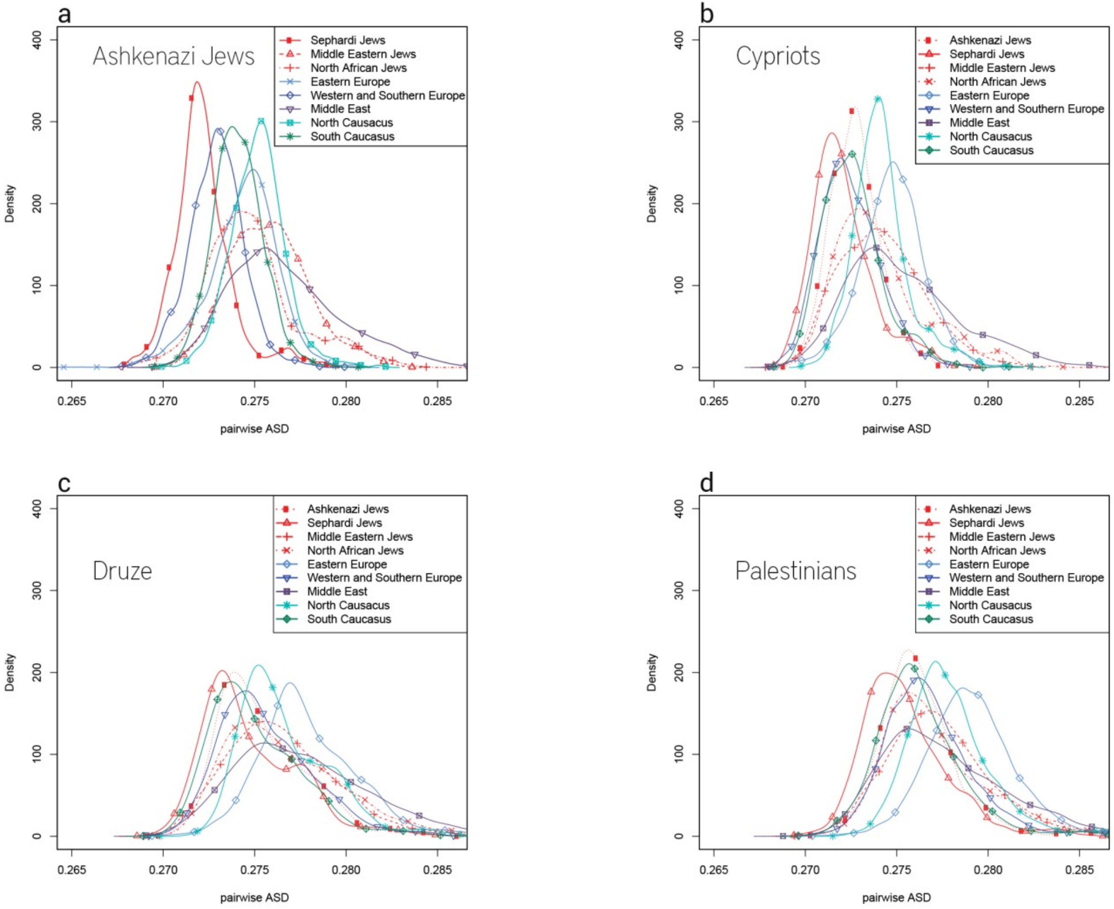
Figure 6: Average identity-by-descent sharing between different population groups and Ashkenazi Jews.

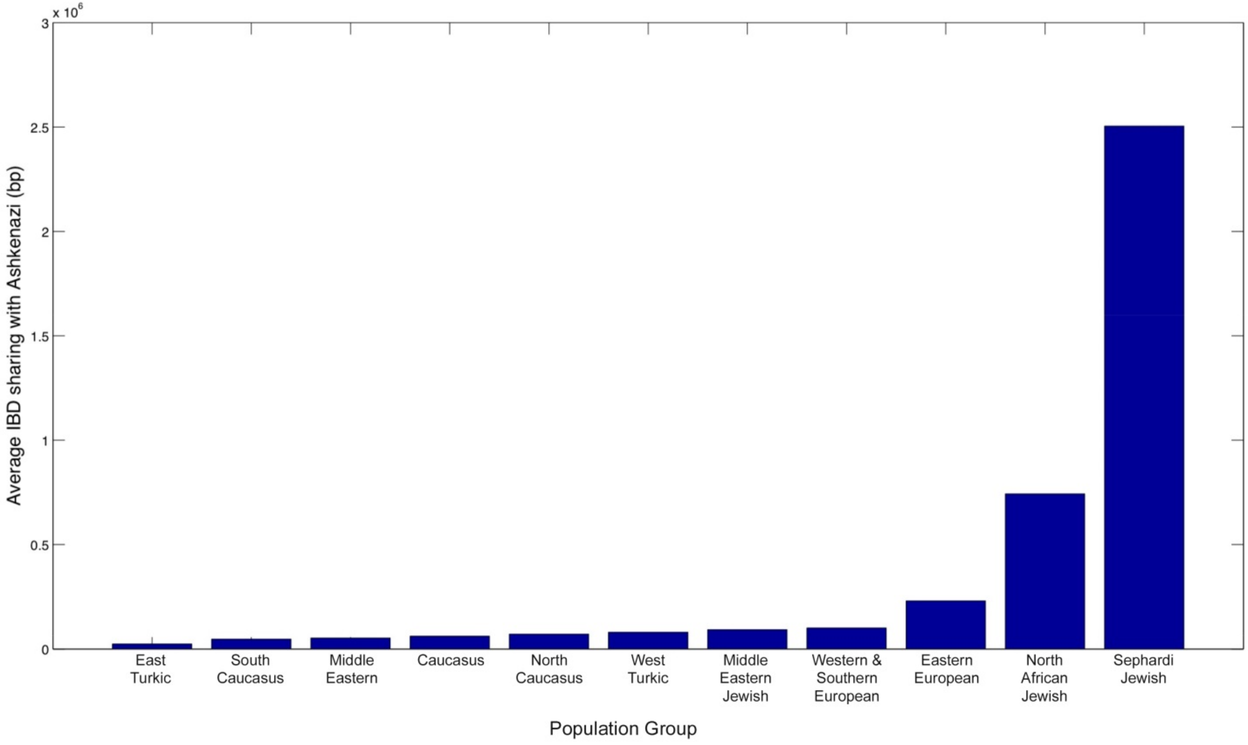

Chu
Tat
Rus
Mrd
Rus
Rus
Rus
Ukr
750 C.E.
850 C.E.
Mol
600 C.E.
Atil
Khazaria
Ady
Nog
Kab
Kum
Abk
Blk
OsN
Che
Tab
Lez
GeoJ
Geo
AzeJ
Arm
Tur
Aze
Kur
Cyp
Syr
SyrJ
KurJ
Irn
Leb
Sam
Pal
Dru
IrqJ
IrnJ
Bed
Jor
Egy
Orc
Swe
Est
Lit
Bel
AshJ
Pol
Ukr
Fre
Hun
Mol
Rom
Bas
ItB
Cro
Bul
SepJ
Spa
ItT
ItA
Sar
ItaJ
Gre
ItS
Mor
MorJ
Moz
TunJ
AlgJ
LibJ
Sau
Yem
YemJ
Man
Eth
EthJ
Yor
Bia
Mbu
Ban
San
Ban
UzbJ
Uzb
Tkm
Taj
Kyr
Uyg
Alt
Tuv
Brt
Mon
Yak
Haz
Kal
Bur
Pat
Bal
Bra
Mak
Sin
Guj
MumJ
KaN
Hal
CocJ
Mal
Sak
Pan
Cam
Jap
Han


A
B
Africa
Arabia
South-Caucasus
North-Caucasus
Central-Asia
Western-Europe
Eeastern-Europe
Jewish
Near-East
Siberia
East-Asia
South-Asia-(Pakistan)
South-Asia-(India)
PC2: 2.68%
0.06
0.04
0.02
0.00
-0.02
PC1: 3.55%
0.08
0.06
0.04
0.02
0.00
-0.02
Latitude
50
45
40
35
30
25
20
Longitude
20
40
60
80
100
120


San
Biaka Pygmies
Mbuti Pygmies
Bantu
Mandenka
Yoruba
Ethiopian
Ethiopian Jewish
Mozabite
Moroccan
Egyptian
Saudi
Yemeni
Yemenite Jewish
Bedouin
Palestinian
Jordanian
Syrian
Lebanese
Druze
Samaritan
Cypriot
Moroccan
Algerian
Tunisian
Libyan
Bulgarian
Turkish
§
Italian
Western
Central/Eastern
Georgian
Azerbaijani
Iranian
¶
Kurdish
Iraqi
Uzbekistani
Turkish
Kurd
Iranian
Spanish
French
French Basque
Sardinian
Sicilian
Bergamo
Tuscan
Abruzzo
Italian
Greek
Croat
Bulgarian
Hungarian
Romanian
Moldavian
Orcadian
Swedish
Polish
Ukranian
Belarusian
Voronez
Kursk
Orjol
Smolensk
Russian
Lithuanian
Estonian
Mordovian
Chuvash
Tatar
Armenian
Azeri
Georgian
Abkhasian
Balkar
Kabardin
Adygei
North Ossetian
Chechen
Lezgin
Tabasaran
Kumyk
Nogai
Turkmen
Tajik
Uzbek
Kyrgyz
Uygur
Altaian
Hazara
Kalash
Pathan
Burusho
Balochi
Brahui
Makrani
Sindhi
Gujarati
Mumbai Jewish
1
North Kannadi
234
Cochin Jewish
1 Halakipikki
2 Malayan
3 Paniya
4 Sakilli
India
Cambodian
Han
Mongolian
Japanese
Buryat
Tuvinian
Yakut
Sephardi
Ashkenazi
Middle Eastern
Africa
Middle East
Jewish
Europe
Caucasus
Central Asia*
South Asia
East Asia


Algerian_Jewish
Eastern_Ashkenazi_Jewish
Western_Ashkenazi_Jewish
Azerbaijani_Jewish
Cochin_Jewish
Ethiopian_Jewish
Georgian_Jewish
Iranian_Jewish
Iraqi_Jewish
Italian_Jewish
Kurdish_Jewish
Libyan_Jewish
Moroccan_Jewish
Mumbai_Jewish
Sephardic_Jewish_Belmonte
Sephardic_Jewish_Bulgaria
Sephardic_Jewish_Turkey
Syrian_Jewish
Tunisian_Jewish
Uzbekistani_Jewish
Yemenite_Jewish
1.00
0.95
0.90
0.85
0.80
Algerian_Jewish
Moroccan_Jewish
Tunisian_Jewish
Italian_Jewish
Sephardic_Jewish_Bulga
Sephardic_Jewish_Turke
Libyan_Jewish
Western_Ashkenazi_Je
Syrian_Jewish
Sicilian
Cypriot
Eastern_Ashkenazi
Sephardic_Jewish
Samaritan
Abruzzo
Lebanese
Druze
Palestinia
Jordania
Syrian
Eastern_Ashkenazi_Jewis
Sicilian
Abruzzo
Western_Ashkenazi_Jew
Sephardic_Jewish_Bul
Sephardic_Jewish_Be
Italian_Jewish
Tuscan
Greek
Sephardic_Jewish_T
Moroccan_Jewish
Algerian_Jewish
Tunisian_Jewish
Bergamo
Cypriot
Syrian_Jew
Libyan_Jew
Sardin
Bulg
Span
Western_Ashkenazi_Jewi
Sicilian
Sephardic_Jewish_Bulga
Italian_Jewish
Eastern_Ashkenazi_Jew
Moroccan_Jewish
Sephardic_Jewish_Turk
Sephardic_Jewish_Bel
Abruzzo
Algerian_Jewish
Tunisian_Jewish
Tuscan
Libyan_Jewish
Cypriot
Greek
Syrian_Jewish
Sardinian
Bergamo
Dru
Leb
Azerbaijani_Jewish
Georgian_Jewish
Iranian_Jewish
Armenian
Kurdish_Jewish
Uzbekistani_Jewish
Iraqi_Jewish
Azeri
Kurd
Druze
Iranian
Syrian
Turkish
Georgian
Lebanese
Abkhasian
Cypriot
Syrian_Jewish
Samaritan
Jordanian
Cochin_Jewish
Gujarati
Mumbai_Jewish
Sakilli
Halakipiki
North_Kannadi
Paniya
Malayan
Sindhi
Burusho
Pathan
Ethiopian_Jewish
Ethiopian
Georgian_Jewish
Azerbaijani_Jewish
Iranian_Jewish
Iraqi_Jewish
Kurdish_Jewish
Armenian
Uzbekistani_Jewish
Druze
Syrian
Azeri
Kurd
Lebanese
Iranian
Turkish
Syrian_Jewish
Cypriot
Samaritan
Georgian
Jordanian
Abkhasian
Iranian_Jewish
Kurdish_Jewish
Azerbaijani_Jewish
Iraqi_Jewish
Georgian_Jewish
Uzbekistani_Jewish
Armenian
Iranian
Kurd
Syrian
Azeri
Druze
Lebanese
Turkish
Georgian
Samaritan
Syrian_Jewish
Cypriot
Abkhasian
Jordanian
Iraqi_Jewish
Kurdish_Jewish
Iranian_Jewish
Georgian_Jewish
Azerbaijani_Jewish
Syrian
Druze
Uzbekistani_Jewish
Armenian
Lebanese
Iranian
Samaritan
Kurd
Azeri
Syrian_Jewish
Jordanian
Cypriot
Palestinian
Turkish
Georgian
Italian_Jewish
Sephardic_Jewish_Bulgar
Sephardic_Jewish_Turkey
Moroccan_Jewish
Algerian_Jewish
Western_Ashkenazi_Jew
Tunisian_Jewish
Sicilian
Eastern_Ashkenazi_Je
Libyan_Jewish
Cypriot
Syrian_Jewish
Sephardic_Jewish_
Abruzzo
Lebanese
Druze
Samaritan
Tuscan
Greek
Pales
Kurdish_Jewish
Iranian_Jewish
Iraqi_Jewish
Georgian_Jewish
Azerbaijani_Jewish
Uzbekistani_Jewish
Armenian
Syrian
Iranian
Druze
Kurd
Azeri
Lebanese
Turkish
Samaritan
Syrian_Jewish
Jordanian
Cypriot
Georgian
Palestinian
Libyan_Jewish
Tunisian_Jewish
Algerian_Jewish
Moroccan_Jewish
Sephardic_Jewish_Tur
Italian_Jewish
Sephardic_Jewish_Bu
Syrian_Jewish
Cypriot
Samaritan
Western_Ashkena
Lebanese
Druze
Palestinian
Sicilian
Jordanian
Eastern_As
Syrian
Sephardic
Iraqi
Moroccan_Jewish
Algerian_Jewish
Tunisian_Jewish
Italian_Jewish
Sephardic_Jewish_Bulga
Sephardic_Jewish_Turke
Libyan_Jewish
Western_Ashkenazi_Je
Sicilian
Syrian_Jewish
Cypriot
Eastern_Ashkenazi
Sephardic_Jewish
Abruzzo
Samaritan
Lebanese
Druze
Palestin
Tuscan
Jordan
Mumbai_Jewish
Cochin_Jewish
Gujarati
Sindhi
Sakilli
Burusho
Halakipiki
Pathan
North_K
Paniy
Mal
Bal
B
Sephardic_Jewish_Belmo
Western_Ashkenazi_Je
Sicilian
Eastern_Ashkenazi_Je
Abruzzo
Tuscan
Italian_Jewish
Sephardic_Jewish_
Moroccan_Jewish
Greek
Sardinian
Algerian_Jewish
Bergamo
Sephardic_Jew
Tunisian_Jew
Spanish
Libyan
Basq
Cyp
Sy
Sephardic_Jewish_Bulgari
Italian_Jewish
Sephardic_Jewish_Turkey
Moroccan_Jewish
Western_Ashkenazi_Jew
Algerian_Jewish
Tunisian_Jewish
Sicilian
Eastern_Ashkenazi_Je
Libyan_Jewish
Cypriot
Syrian_Jewish
Sephardic_Jewish_
Abruzzo
Tuscan
Lebanese
Druze
Greek
Samaritan
Syria
Sephardic_Jewish_Turkey
Sephardic_Jewish_Bulgar
Italian_Jewish
Algerian_Jewish
Tunisian_Jewish
Moroccan_Jewish
Cypriot
Libyan_Jewish
Western_Ashkenazi_Je
Syrian_Jewish
Sicilian
Eastern_Ashkenazi
Lebanese
Abruzzo
Druze
Sephardic_Jewi
Samaritan
Syrian
Palestinian
Jordanian
Syrian_Jewish
Cypriot
Druze
Lebanese
Sephardic_Jewish_Tur
Tunisian_Jewish
Samaritan
Libyan_Jewish
Algerian_Jewish
Sephardic_Jewish_B
Italian_Jewish
Syrian
Moroccan_Jewish
Georgian_Jewish
Iraqi_Jewish
Palestinian
Jordanian
Azerbaijani_Jew
Kurdish_Jewish
Armenian
Tunisian_Jewish
Libyan_Jewish
Moroccan_Jewish
Algerian_Jewish
Italian_Jewish
Sephardic_Jewish_Turke
Sephardic_Jewish_Bulg
Syrian_Jewish
Cypriot
Western_Ashkenazi
Sicilian
Eastern_Ashkena
Samaritan
Lebanese
Druze
Sephardic_Jew
Palestinian
Jordanian
Abruzzo
Syrian
Uzbekistani_Jewish
Iranian_Jewish
Kurdish_Jewish
Kurd
Azerbaijani_Jewish
Iranian
Armenian
Azeri
Georgian_Jewish
Iraqi_Jewish
Turkish
Syrian
Georgian
Druze
Abkhasian
Lebanese
Cypriot
Syrian_Jewis
North_Osseti
Adygei
Yemenite_Jewish
Saudi
Bedouin
Yemeni
Egyptian
Palestinian
Jordania
Sa


a
Ashkenazi Jews
Sephardi Jews
Middle Eastern Jews
North African Jews
Eastern Europe
Western and Southern Europe
Middle East
North Causacus
South Caucasus
Density
0
100
200
300
400
0.265
0.270
0.275
0.280
0.285
pairwise ASD
b
Cypriots
Ashkenazi Jews
Sephardi Jews
Middle Eastern Jews
North African Jews
Eastern Europe
Western and Southern Europe
Middle East
North Causacus
South Caucasus
Density
0
100
200
300
400
0.265
0.270
0.275
0.280
0.285
pairwise ASD
c
Druze
Ashkenazi Jews
Sephardi Jews
Middle Eastern Jews
North African Jews
Eastern Europe
Western and Southern Europe
Middle East
North Causacus
South Caucasus
Density
0
100
200
300
400
0.265
0.270
0.275
0.280
0.285
pairwise ASD
d
Palestinians
Ashkenazi Jews
Sephardi Jews
Middle Eastern Jews
North African Jews
Eastern Europe
Western and Southern Europe
Middle East
North Causacus
South Caucasus
Density
0
100
200
300
400
0.265
0.270
0.275
0.280
0.285
pairwise ASD


x 10^6
3
2.5
2
1.5
1
0.5
0
Average IBD sharing with Ashkenazi (bp)
East Turkic
South Caucasus
Middle Eastern
Caucasus
North Caucasus
West Turkic
Middle Eastern Jewish
Western & Southern European
Eastern European
North African Jewish
Sephardi Jewish
Population Group

Supplemental Material

Supplemental File 1: List of samples and populations used in this study. The file details the population name, region, 3-letter code identifier, color, geographic coordinates, and source for the first publication.

Supplemental Table 1: *P*-values for the two-dimensional bootstrap approach to determine statistical significance of average allele-sharing distances.

Supplemental Table 2: *P*-values for tests of different levels of identity-by-descent sharing of population groups with Ashkenazi Jews.

Supplemental Figure 1: Scatter plot of the first and second principal components for all samples included in the study.

Supplemental Figure 2: Scatter plot of the first and third principal components for all samples included in the study.

Supplemental Figure 3: ADMIXTURE plots showing results at *K*=2 to 11, *K*=14, *K*=15, and *K*=17 (see methods for choice of *K*). Western Ashkenazi Jews: France, Germany, Netherlands; Central and Eastern Ashkenazi Jews: Austria, Belorussia, Hungary, Latvia, Lithuania, Poland, Romania, Russia.

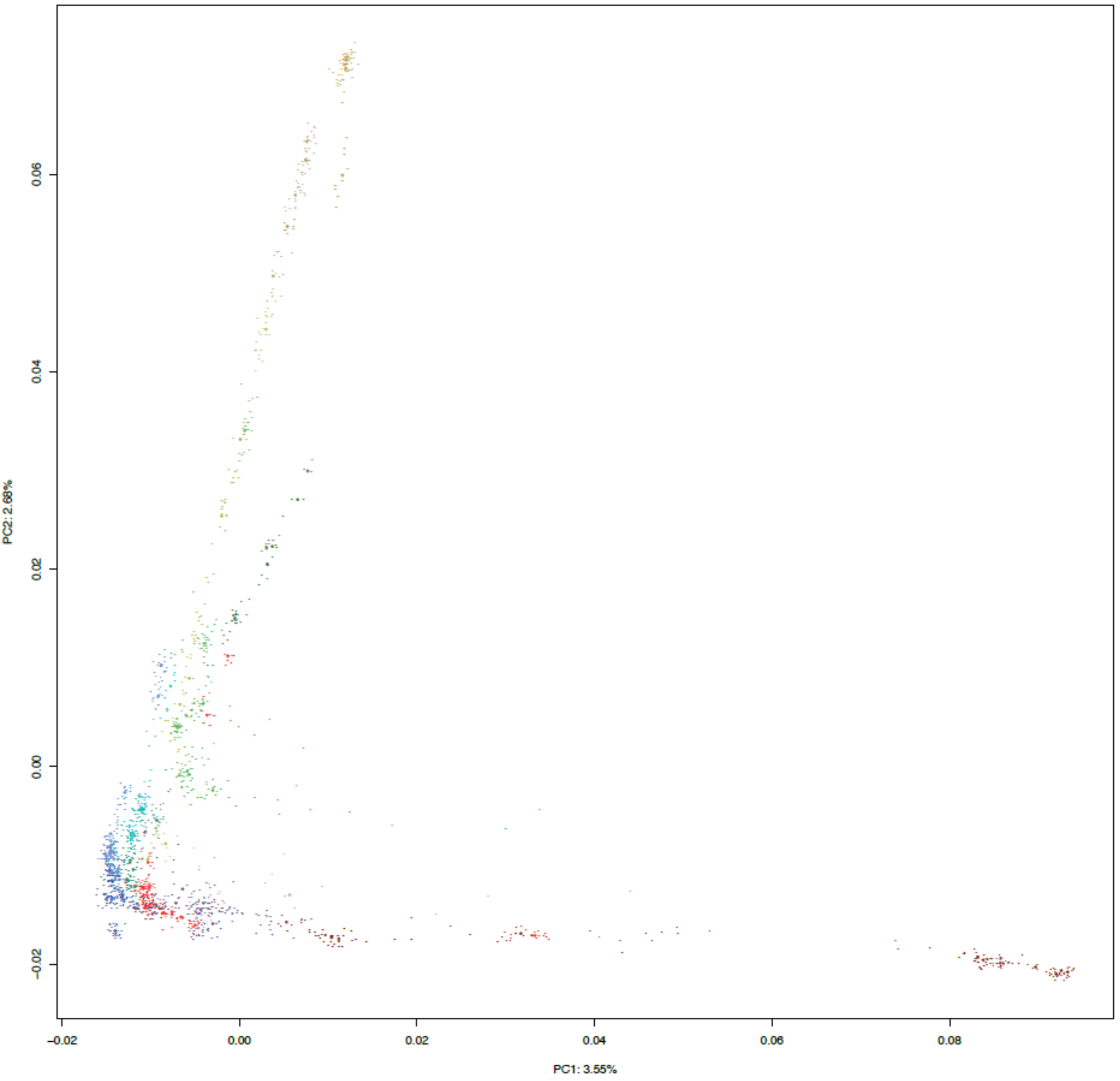
Supplemental Figure 4: Cross-validation errors of the ADMIXTURE runs at *K* values 2 to 20 (with magnification for *K*=6 to *K*=14).

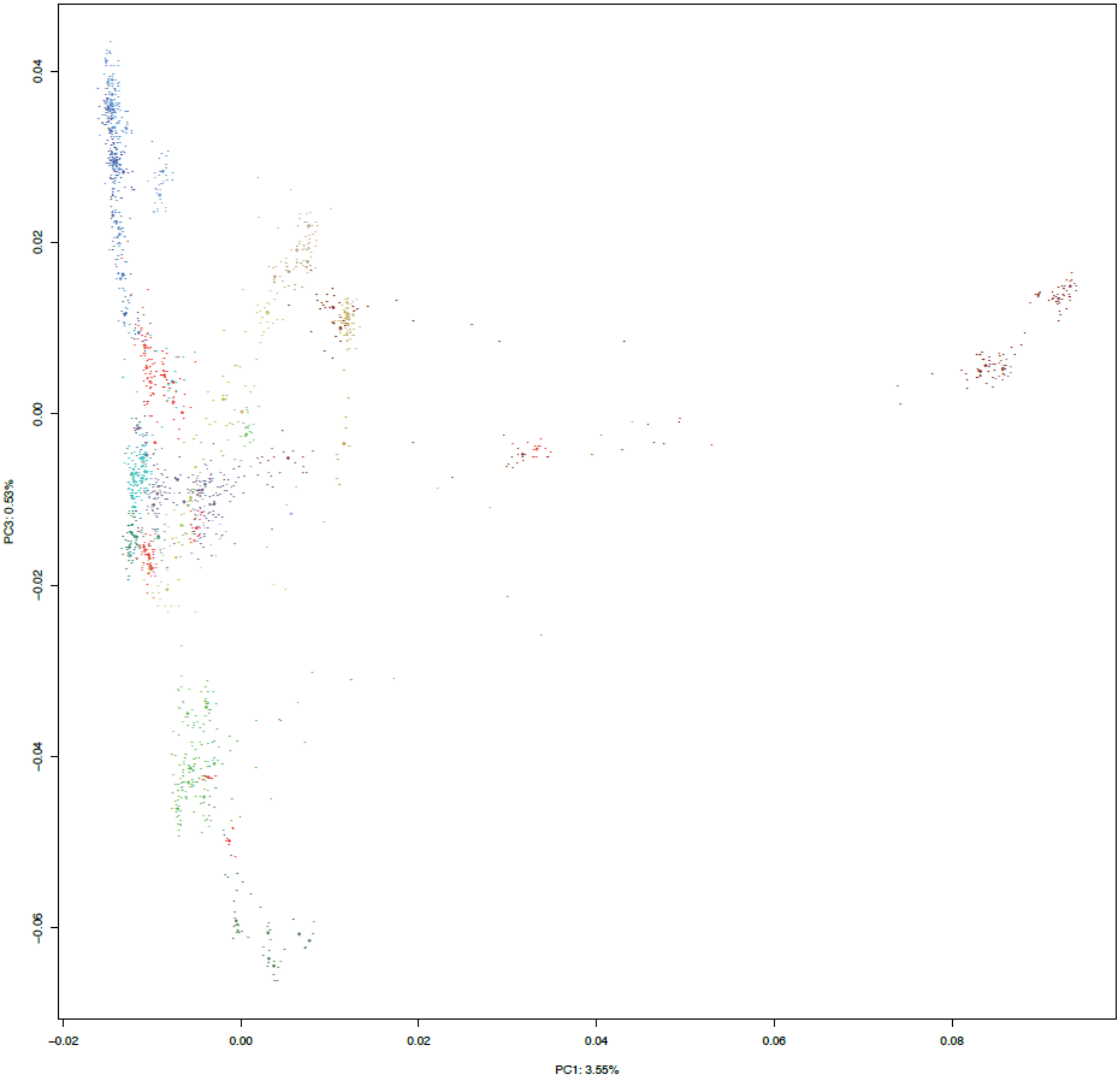
Supplemental Figure 5: Maximum difference in log likelihood (LL) scores in fractions (0.05, 0.1, 0.2) of runs with the highest LL scores.

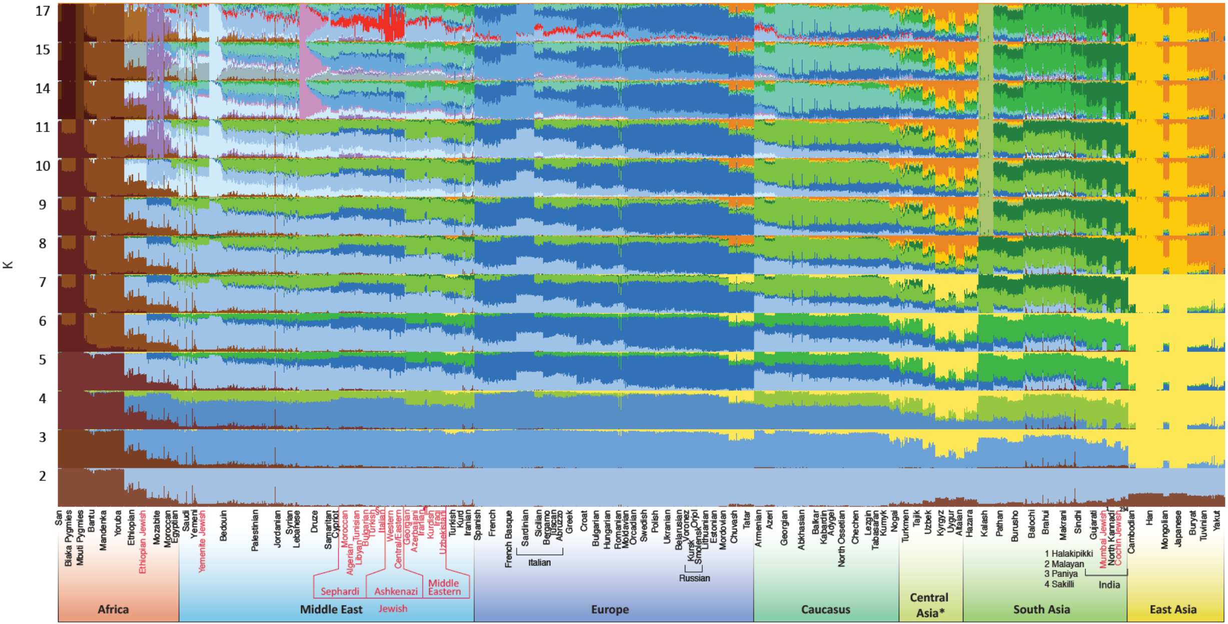
Supplemental Figure 6: CLUMPP scores vs. log-likelihood (LL) differences. The x axis shows LL differences to the run arriving at the highest LL score as a function of *K*. Provided the LL difference is <100, the CLUMPP score is 0.9999 or greater. The highlighted exceptions (red circle) come from *K*=17, where 10 runs display low LL differences from the best run (3.5 to 5.5 LL units) and at the same time show a CLUMPP score well below 1. Note, however, that at *K*=17, 11 runs reach a CLUMPP score of ~1 and are within 2 LL units from the best run.

Supplemental Figure 7: CLUMPP scores vs. log-likelihood differences for different values of *K*. The x axis shows all runs, sorted by LL-score difference to the run arriving at the highest LL score among the runs at a particular *K*. Similar LL scores translate to similar CLUMPP scores. However, no direct relation exists between LL scores and CLUMPP scores; an increase in the LL scores difference can result in higher CLUMPP scores (see *K*=7 in panel A). However, if the CLUMPP score is ~1, then the LL difference is <20. Exceptions in this respect at *K*=17 (see **Supplemental Figure 6**) are highlighted. Panel B has a zoomed y axis to highlight LL score differences <20.

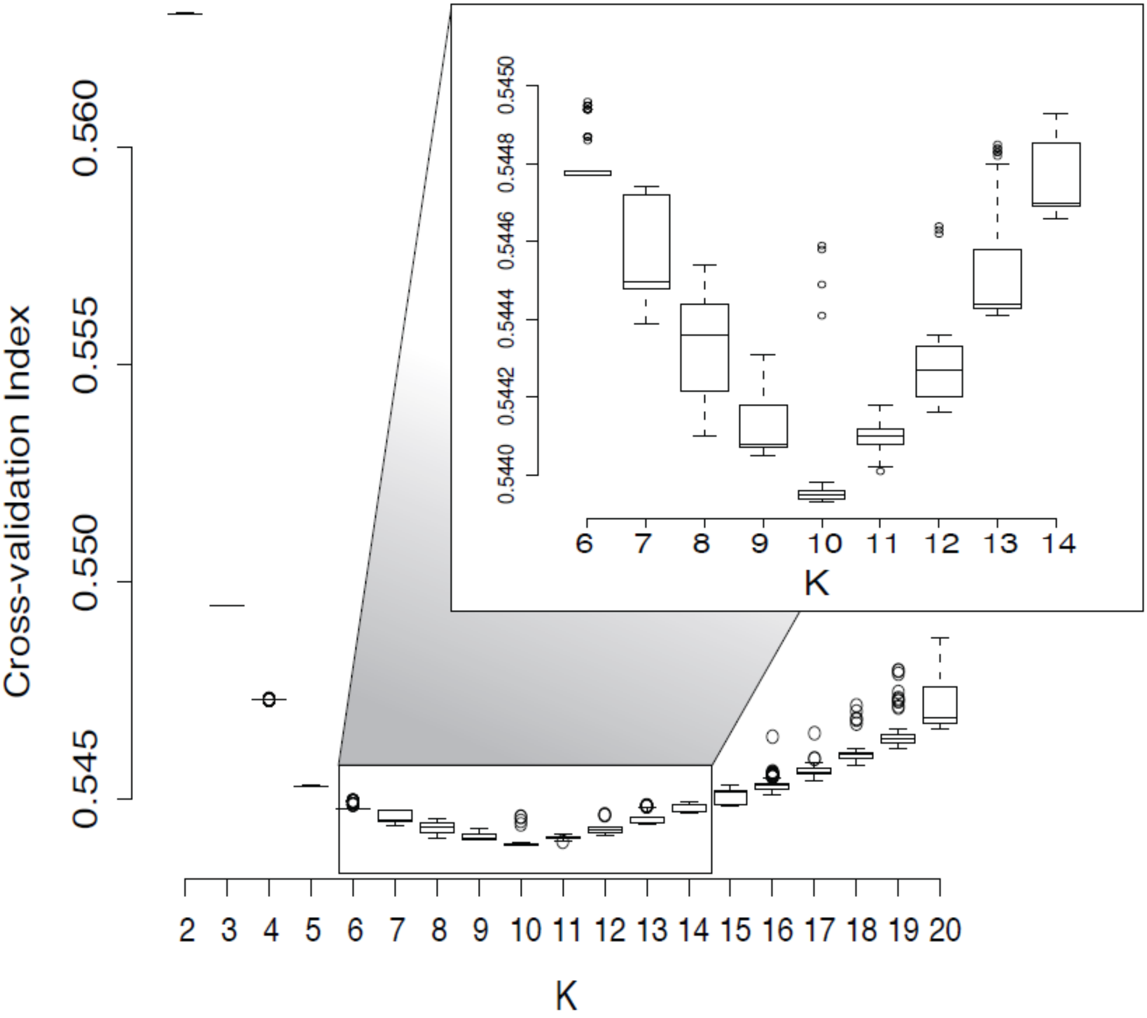
Supplemental Figure 8: Density plot of the pairwise allele-sharing distances between individuals from a specific population and individuals from a population group**.** a. Eastern Ashkenazi Jews. b. Western Ashkenazi Jews.

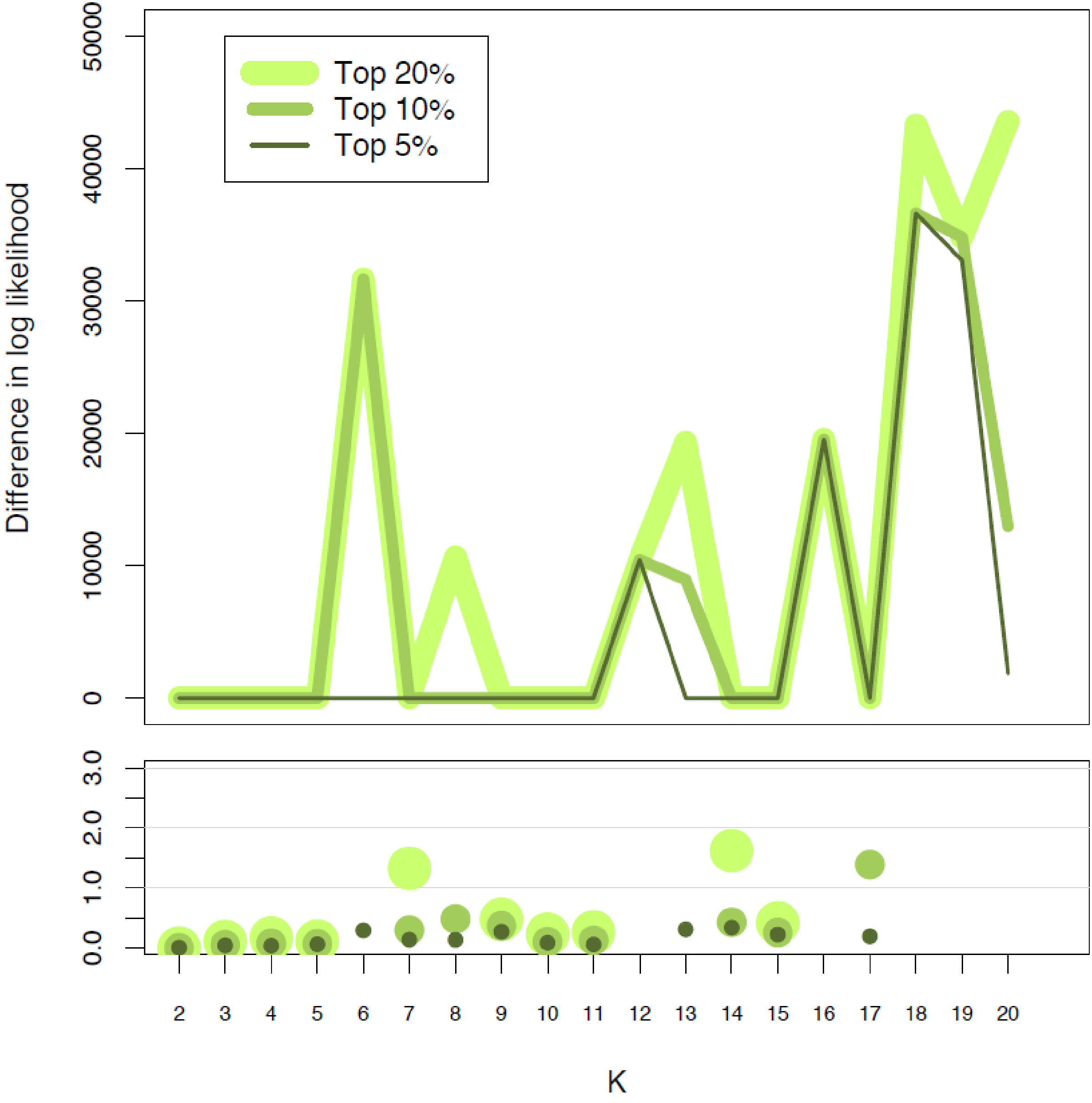

PC2: 2.68%
0.06
0.04
0.02
0.00
−0.02
−0.02
0.00
0.02
0.04
0.06
0.08
PC1: 3.55%

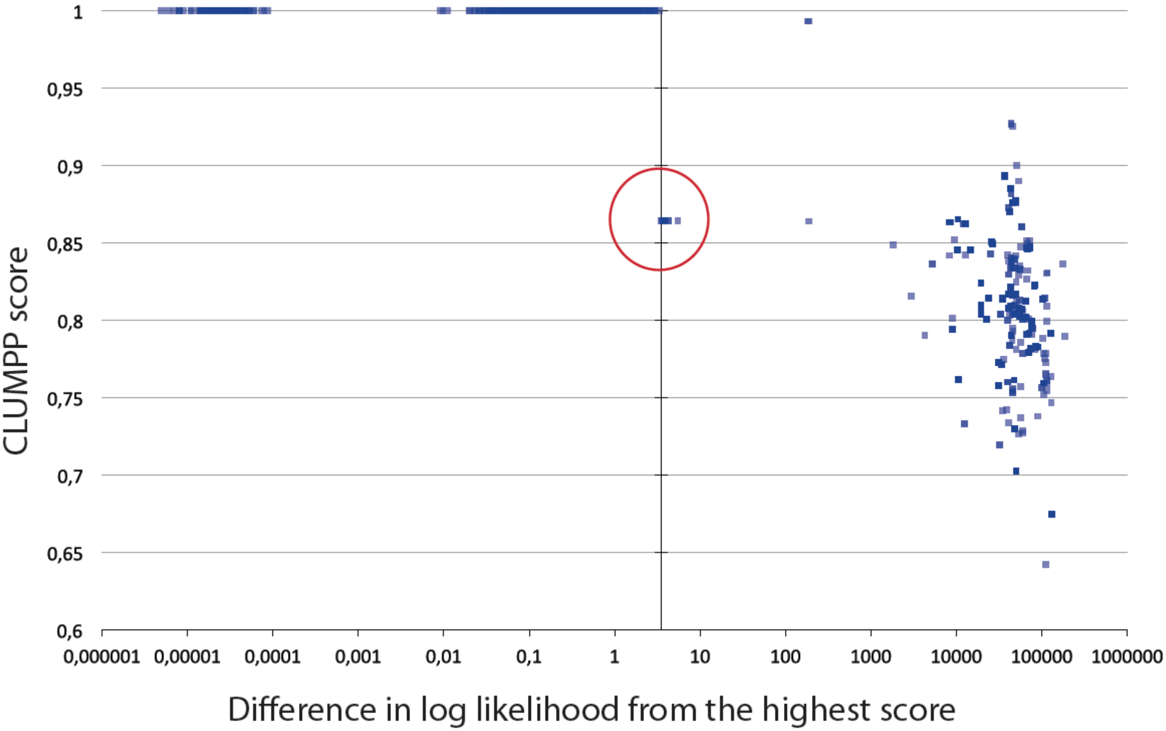

0.04
0.02
0.00
−0.02
−0.04
−0.06
PC3: 0.53%
−0.02
0.00
0.02
0.04
0.06
0.08
PC1: 3.55%

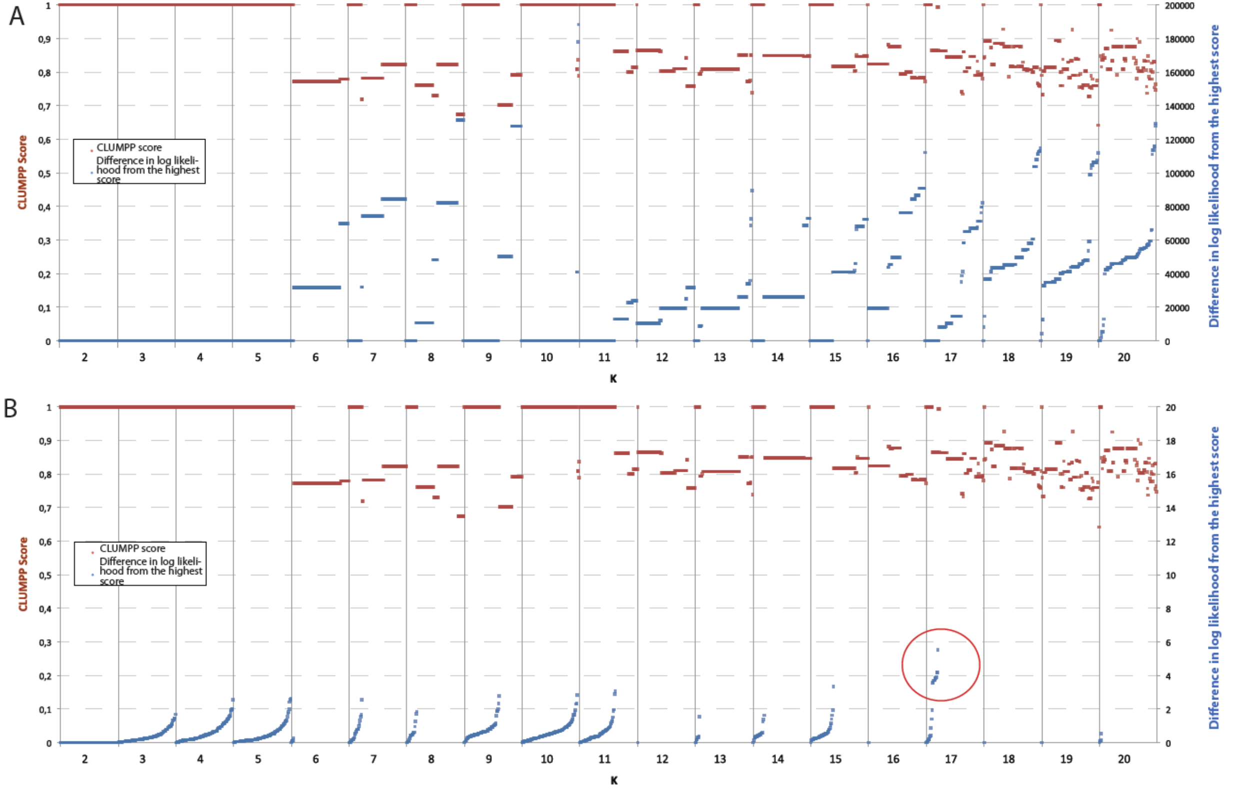


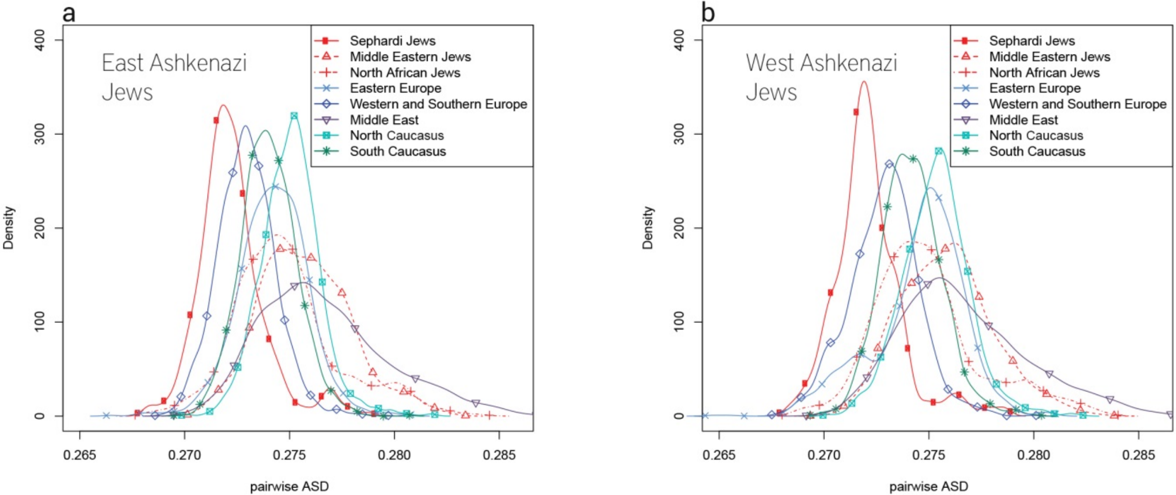
K
17
15
14
11
10
9
8
7
6
5
4
3
2
San
Biaka Pygmies
Mbuti Pygmies
Bantu
Mandenka
Yoruba
Ethiopian
Ethiopian Jewish
Mozabite
Moroccan
Egyptian
Saudi
Yemeni
Yemenite Jewish
Bedouin
Palestinian
Jordanian
Syrian
Lebanese
Druze
Samaritan
Cypriot
Moroccan
Algerian
Tunisian
Libyan
Bulgarian
Turkish
Italian
Western
Central/Eastern
Georgian
Azerbaijani
Iranian
Kurdish
Iraqi
Uzbekistani
Turkish
Kurd
Iranian
Sephardi
Ashkenazi
Middle Eastern
Spanish
French
French Basque
Sardinian
Sicilian
Bergamo
Tuscan
Abruzzo
Italian
Greek
Croat
Bulgarian
Hungarian
Romanian
Moldavian
Orcadian
Swedish
Polish
Ukranian
Belarusian
Voronez
Kursk
Smolensk
Orjol
Russian
Lithuanian
Estonian
Mordovian
Chuvash
Tatar
Armenian
Azeri
Georgian
Abkhasian
Balkar
Kabardin
Adygei
North Ossetian
Chechen
Lezgin
Tabasaran
Kumyk
Nogai
Turkmen
Tajik
Uzbek
Kyrgyz
Uygur
Altaian
Hazara
Kalash
Pathan
Burusho
Balochi
Brahui
Makrani
Sindhi
Gujarati
Mumbai Jewish
1
North Kannadi
Cochin Jewish
234
Cambodian
Han
Mongolian
Japanese
Buryat
Tuvinian
Yakut
1 Halakipikki
2 Malayan
3 Paniya
4 Sakilli
India
Africa
Middle East
Jewish
Europe
Caucasus
Central Asia*
South Asia
East Asia


Cross-validation Index
0.545
0.550
0.555
0.560
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
K
0.5440
0.5442
0.5444
0.5446
0.5448
0.5450
6
7
8
9
10
11
12
13
14
K


Top 20%
Top 10%
Top 5%
Difference in log likelihood
0
10000
20000
30000
40000
50000
0.0
1.0
2.0
3.0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
K


1
0,95
0,9
0,85
0,8
0,75
0,7
0,65
0,6
CLUMPP score
0,000001
0,00001
0,0001
0,001
0,01
0,1
1
10
100
1000
10000
100000
1000000
Difference in log likelihood from the highest score


A
CLUMPP Score
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
CLUMPP score
Difference in log likeli-hood from the highest score
Difference in log likelihood from the highest score
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
K
B
CLUMPP Score
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
CLUMPP score
Difference in log likeli-hood from the highest score
Difference in log likelihood from the highest score
20
18
16
14
12
10
8
6
4
2
0
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
K


a
East Ashkenazi Jews
b
West Ashkenazi Jews
Sephardi Jews
Middle Eastern Jews
North African Jews
Eastern Europe
Western and Southern Europe
Middle East
North Caucasus
South Caucasus
Density
0
100
200
300
400
0.265
0.270
0.275
0.280
0.285
pairwise ASD