Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry

G. Atzmon[1,4*], L. Hao[2*+], I. Pe'er[3*], C. Velez[2], A. Pearlman[2], P.F. Palamara[3], B. Morrow[4], E. Friedman[5], C. Oddoux[2], E. Burns[1], H. Ostrer[2#]

1) Department of Medicine, Albert Einstein College of Medicine, Bronx, NY

2) Human Genetics Program, Department of Pediatrics, New York University School of Medicine, New York, NY

3) Department of Computer Science, Columbia University, New York, NY

4) Department of Genetics, Albert Einstein College of Medicine, Bronx, NY

5) The Susanne Levy Gertner Oncogenetics Unit, the Danek Gertner Institute of Human Genetics, Chaim Sheba Medical Center, 52621, Tel-Hashomer, and the Sackler School of Medicine, Tel-Aviv University, Tel-Aviv Israel

[#]Corresponding author: Harry Ostrer, M.D.
Human Genetics Program
Department of Pediatrics
New York University School of Medicine
550 First Avenue, MSB136
New York, NY 10016
tel: 212 263-7596
fax: 212 263-3477
email: harry.ostrer@nyumc.org

[*]contributed equally to the work

[+]Current address: Center for Genome Informatics, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, NJ

Summary

For more than a century, Jews and non-Jews alike have tried to define the relatedness of contemporary Jewish people. Previous genetic studies of blood group and serum markers suggested that Jewish groups had Middle Eastern origin with greater genetic similarity between paired Jewish populations. However, these and successor studies of monoallelic Y chromosomal and mitochondrial genetic markers did not resolve the issues of within and between-group Jewish genetic identity. Here, genome wide analysis of seven Jewish groups (Iranian, Iraqi, Syrian, Italian, Turkish, Greek and Ashkenazi) and comparison with non-Jewish groups demonstrated distinctive Jewish population clusters, each with shared Middle Eastern ancestry, proximity to contemporary Middle Eastern populations and variable degrees of European and North African admixture. Two major groups were identified by principal component, phylogenetic, and identity by descent (IBD) analysis -- Middle Eastern Jews and European/Syrian Jews. The IBD segment sharing and the proximity of European Jews to each other and to southern European populations suggested similar origins for European Jewry and refuted large-scale genetic contributions of Central and Eastern European and Slavic populations to the formation of Ashkenazi Jewry. Rapid decay of IBD in Ashkenazi Jewish genomes was consistent with a severe bottleneck followed by large expansion, such as occurred with the so-called *demographic miracle* of population expansion from 50,000 people at the beginning of the 15th century to 5,000,000 people at the beginning of the 19th century. Thus, this study demonstrates that European/Syrian and Middle Eastern Jews represent a series of geographical isolates or clusters woven together by shared IBD genetic threads.

Introduction

Jews originated as a national and religious group in the Middle East during the second millennium B.C.E.,[1] and have maintained continuous genetic, cultural, and religious traditions since that time, despite a series of Diasporas.[2] Middle Eastern (Iranian and Iraqi) Jews date from communities that were formed in the Babylon and Persian Empires in the fourth to sixth centuries B.C.E.[3-4] Jewish communities in the Balkans, Italy, North Africa, and Syria were formed during Classical Antiquity and then admixed with Sephardic Jews who migrated following their expulsion from the Iberian Peninsula in the late 15th century.[5] Ashkenazi Jews are thought to have settled in the Rhine Valley during the first millennium of the Common Era, then migrated into Eastern Europe between the 11th to 15th centuries, although alternative theories involving descent from Sorbs (Slavic speakers in Germany) and Khazars have also been proposed.[6-7] Admixture with surrounding populations had an early role in shaping world Jewry, but, during the past 2000 years, may have been limited by religious law as Judaism evolved from a proselytizing to an inward-looking religion.[8]

Earlier genetic studies on blood groups and serum markers suggested that Jewish Diaspora populations had Middle Eastern origin, with greater genetic similarity between paired Jewish populations than with non-Jewish populations.[9-11] These studies differed in their interpretation of the degree of admixture with local populations. Recent studies of Y chromosomal and mitochondrial DNA haplotypes have pointed to founder effects of both Middle Eastern and local origin, yet, the issue of how to characterize Jewish people as mere co-religionists or as genetic isolates that may be closely or loosely related remains unresolved.[12-16] To improve the understanding about the relatedness of contemporary Jewish groups, genome

wide analysis and comparison with neighboring populations was performed for representatives of 3 major groups of the Jewish Diaspora -- Eastern European Ashkenazim, Italian, Greek and Turkish Sephardim, and Iranian, Iraqi, and Syrian Mizrahim (Middle Easterners).

Material and Methods

Recruitment and genotyping of Jewish populations. Participants were recruited from the Iranian, Iraqi, Syrian and Ashkenazi Jewish communities in the metropolitan New York region. Participants were recruited from the Turkish Sephardic Jewish community in Seattle, from the Greek Sephardic Jewish communities in Thessaloniki and Athens and from the Italian Jewish community in Rome, the latter as previously described.[17] All of the recruitments took place following a New York University School of Medicine Institutional Review Board-approved protocol (07-333 "Origins and Migrations of Jewish People"). Additional recruitment of Iraqi and Turkish Sephardic Jews occurred at Sheba Medical Centre in Tel Hashomer, Israel following a local ethics committee and an Israeli Ministry of Health Institutional Review Board approved protocol. In every case, subjects provided informed consent. They were included only if all 4 grandparents came from the same Jewish community. Subjects were excluded if they were known first or second degree relatives of other participants or were found to have IBD coefficients $\geq .30$ by analysis of microarray data. Genotyping was performed using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affy v 6) at the genomic facility at Albert Einstein College of Medicine.

DNA preparation for SNP array analysis. DNA was prepared according to standard methods. Quality and quantity of genomic DNA was determined by agarose gel electrophoresis to assure that only high molecule weight DNA was present and by absorbance at 230, 260, and

280 nm to determine DNA concentration and assure that protein and organic contaminants were not present.

Genotyping. Genomic DNA samples were genotyped with the Affy v 6 in accordance with the manufacturer's protocols. A total of 305 Jewish samples were successfully genotyped at call rates > 99% and with no gender mismatch. The resulting individuals were tested for relatedness using genome wide IBD estimates. Samples were excluded if the IBD coefficients were >.30, as this suggests hidden relatedness. To assure that members came from the stated community, the SMART PCA program from EIGENSOFT package used to remove genetic outliers (defined as having greater than six standard deviations from the mean PC position on at least one of the top ten eigenvectors). A total of 14 were observed and these samples were removed. Ultimately, 237 samples were used for comparative analyses. Results of samples of known European origin that were run on Affy v 6 arrays were included in the PCA analysis. These overlapped completely with the results from the current study, indicating the absence of a batch effect.

Reference populations. *HGDP dataset* – The Jewish dataset was analyzed along with a selected HGDP datasets. The original HGDP dataset had 1043 unrelated individuals from 52 world-wide populations.[18] First, 28 extreme outliers identified by three independent preliminary PCA runs on a set of small randomly selected SNPs was removed. To reduce the size of the dataset, members of related population groups were combined, including Pakistani (Balochi, Brahui, Burusho, Makrani, Pastun, Sindi, Uyghur), Southern American (Colombian, Karitiana, Maya, Pima, Surui), Central/Southern African (Bantu, Biaka, Mandenka, Mbuti Pygmy, Mozabite, San, Yoruba), and East Asian (Khmer, Dai, Daur, Northern Chinese Han, Southern

Chinese Han, Hezhen, Japanese, Lahu, Miao, Mongolian, Naxi, Oroqen, She, Tu, Tujia, Xibo, Yi), and then 25 samples were selected randomly from each population. The final number of samples in this selected HGDP dataset was 418. These came from 16 populations: North African (Mozabite), Central and South African, East Asian, Southern American, Pakistani_Hazara, Pakistani_Kalash, Pakistani_Other, Middle Eastern_Bedouin, Middle Eastern_Druze, Middle Eastern_Palestinian, Adygei, Russian, Basque, French, Northern Italian (Bergamo and Tuscan), and Sardinian . No significant differences were observed in the results when different datasets containing independent, randomly selected samples were used. To get a closer view of Jewish population structure, a localized dataset was generated that combined the Jewish populations, 3 Middle Eastern non-Jewish populations, and 6 European populations. The HGDP samples were genotyped on the Illumina HumanHap650K Beadchips, as previously described. After filtering SNPs with call rate <95% and extracting overlapping SNP sets between two different platforms, 164,894 SNPs were used for further analysis.

PopRes dataset - The Population Reference Sample (PopRes) project included over 4,000 individuals of African-American, East Asian, South Asian, Mexican, and European origin after quality control, of which 2,407 individuals of unmixed ancestry were collected from a wide variety of European countries.[19] These were genotyped on the Affymetrix 500K chip. To study the relationships between Jewish and European population, a localized dataset was generated that combined the Jewish dataset with selected PopRes data. First, 25 extreme outliers identified by three independent preliminary PCA runs on a small set of randomly selected SNPs were removed. Next, each of 2,407 European subjects was assigned into one of 10 groups based on geographic region, *South*-Italy, Swiss-Italian; *Southeast*-Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Kosovo, Macedonia, Romania, Serbia, Slovenia, Yugoslavia; *Southwest*-

Portugal, Spain; *East*-Czech Republic, Hungary; *East-Southeast*-Cyprus, Turkey; *Central*-Austria, Germany, Netherlands, Swiss-German; *West*-Belgium, France, Swiss-French, Switzerland; *North*-Denmark, Norway, Sweden; *Northeast*-Finland, Latvia, Poland, Russia, Ukraine; *Northwest*-Ireland, Scotland, UK. To reduce the size of dataset, 50 samples were randomly selected from each geographic group whenever the sample size was greater than 50. The final number of samples in this selected PopRes dataset was 383. No significant difference in the results was observed when different datasets containing independently selected samples were used. After filtering out SNPs with call rate <95% and extracting overlapping SNP sets between two different platforms, 362,566 SNPs were used for further analysis. This sparser set of SNPs maintains ability to detect IBD, yet is larger than that used in a recent study of genetic structure of the Han Chinese (Fig S9).[19]

Fst, observed heterozygosity, and phylogenetic analysis. Population divergence was measured using the pairwise F_{ST} statistic, calculated with the method of Weir & Cockerham.[22] Confidence intervals of the F_{ST} were calculated by bootstrap resampling, with 500 replications. The genetic diversity across all loci within each population was assessed by using the observed heterozygosity (H_o), calculated from GenePop 4.0 ($1\text{-}Q_{inter}$). The neighbor-joining phylogenetic tree based on pairwise F_{ST} distance was constructed using MEGA4.[23] A Sub-Saharan African population was used as an out-group to root the phylogenetic tree. Bootstrap analyses indicated that the phylogenetic tree is quite robust. Nonetheless, it is unlikely these populations followed a strict tree-like model of evolution, given the abundance of admixture and gene flow between groups.

Principal component and STRUCTURE analysis. Principal component analysis was performed using the *Smartpca* program from the EIGENSOFT package (version 2.0).[24] Except

the initial run (to remove extreme genetic outliers), the analyses were performed without removal of outliers. To infer the population structure, a Bayesian model-based clustering method, implemented in the *STRUCTURE* version 2.2 software package was used for the global dataset.[29] To reduce the running time while still maintaining the information of population structure within the dataset, a subset of 3904 SNPs with highest informativeness across all populations was used. SNPs informativeness was estimated by using average genetic distance difference (δ) among populations studied. For each population pair, δ was calculated as the sum of the absolute differences between allele frequencies. Markers were then ranked and top 5% of SNPs were selected for subsequent *STRUCTURE* analysis. The program was run 10 times for K values 2-6. All structure runs used 30,000 burn-in cycles followed by 30,000 MCMC iterations, assuming correlated allele frequencies and admixture model with separate alpha estimated for each population. The results from all replicates for each K were aligned with CLUMPP.[25] Mean individual Q matrices were plotted using *DISTRUCT*.[26]

Differences between subgroups pairwise *Fst*, IBS and ANOVA. Formal statistical t-testing was performed of each pairwise *Fst* to demonstrate that they differed from zero (tables S2 and S3). Permutation tests were performed for between-group identity-by-state (IBS) with 10,000 permutations for all pair-wise comparisons of 7 Jewish populations. The results showed that comparisons of individuals from the same Jewish populations were genetically much more similar than those from Jewish-non-Jewish and non-Jewish-non-Jewish populations (p values are generally smaller than 10^{-4} -- see table S5 for details). In addition, analysis of variance (ANOVA) was applied to each subgroup's Eigenvalue PCA average to test if paired populations were different.[24]

CNV analysis. Informative CNVs were chosen based on the location of 164,894 SNPs that were used in our SNP analysis. One CNV upstream and one CNV downstream in close proximity to the candidate SNP were included in this analysis. In the cases of SNPs aggregation, CNVs from the regions flanking the SNPs were chosen. PCA analysis was performed on .cel files for 275,000 flanking CNVs from the 237 samples using JMP Genomics 4 (SAS Inc., Cary, NC). To account for biased copy numbers (more than 3 per locus), the values were reassigned new integral values – copy number 1 = 0, copy number 2 = 1, and ≥copy number 3 = 2.

IBD discovery. IBD segments were detected using the GERMLINE algorithm in Genotype Extension.[27] GERMLINE identifies pairwise IBD shared segments in time proportional to the number of individuals processed. Briefly, the algorithm rapidly seeks out short, exact pairwise matches between individuals, and, then, extends from these seeds to long, inexact matches that are indicative of IBD. The output of GERMLINE was used to detect unreported close relatives, who were omitted from the analysis. Two individuals were considered cryptic relatives if their total sharing is larger than 1500 cM and if the average segment length is more than 25cM, suggesting an avuncular or closer relationship. The output was also used to produce sharing densities, sharing graphs and sharing statistics (Appendices - statistical methods).

Inference of population history: To estimate population parameters data was simulated using Genome, with default parameters of recombination rate and block size.[28] Population size and timing of founder/split events were attempted as described in fig. S10 to best fit observed data. Theoretical analysis suggests that the number of IBD segments of a particular length L due to a shared ancestor k generations ago, decreases, for a fixed k, as an exponential function of L. A history of rapid expansion following a recent bottleneck implies that a large fraction of IBD

segments are due to the bottleneck generation, consistent with the exponential decay of shared segments as a function of *L*, that is observed in Ashkenazi samples. In contrast, a fixed-size population will have segments due to ancestors at different generations, producing a different decay pattern, as a sum of exponentials.

Statistical analyses of inter-population differences and neighbor joining trees. To identify whether there are significant genetic differences between Jewish populations, PLINK was used to run permutation test (10,000 permutations) for between-group IBS differences. The neighbor joining tree was generated by using pairwise F_{ST}. To assess the reliability of the NJ tree, SNP loci were randomly sampled 500 times and distance matrix were generated from each sampling dataset. The function "Neighbor" from PHYLIP was used to construct all bootstrap trees, and then "Consense" was used to get bootstrap consensus tree and bootstrap support values for each node.[42]

GERMLINE analysis. *Genotype extension.* In its latest version, GERMLINE can be used in Haplotype Extension or Genotype Extension.[27] The Haplotype Extension is intended to process well-phased data, where it performs with near-perfect accuracy; however, performance can suffer when the data is phased poorly - as can be the case when trio or family data are unavailable. This analysis was performed using the Genotype Extension, where heterozygous markers are treated as wildcards and IBD segments are detected using long segments of mutually homozygous markers. The first stage of GERMLINE searches for seed matches of completely identical haplotypes in computationally phased data. Seeds of k=128 common SNPs were used for Affymetrix SNP 6.0 data, and k=32 common SNPs for the sparser set of SNPs in the intersection this SNP array with the HGDP SNP set. Genotype extension then attempts to extend each seed match between a pair of samples by assuming a Hidden Markov Model (HMM) would

well describe the genotypes of the IBD pair along the IBD segment extending the match. This HMM have been previously used in standard tools, such as PLINK (--segment option), or other work for the entire process of IBD detection.[41] A speedup of the HMM analysis was implemented that advances 64 SNPs at a time, and requires assumptions on genotyping accuracy. In this analysis, one inconsistency was allowed for Affymetrix SNP 6.0 data, and zero inconsistencies for the sparser data.

Filtering regions for informative SNPs. GERMLINE output was filtered to ensure consistency across genotyping platforms and to remove noise by filtering out regions of low information content. SNP density in sliding, non-overlapping blocks across the genome was used to filter shared segments that spanned SNP-sparse regions, particularly, the edges of the centromere and telomere. Specifically, regions that presented less than 100 SNPs per megabase or 100 SNPs per centimorgan were identified and excised and, subsequently, shared segments that were shorter than 3 cM were removed.

Sharing densities. Histograms of post-processed sharing densities were represented by Manhattan-style plots, where the y -axis represents the chance of a random pair of individuals having a shared segment at a SNP: all pairs of individuals sharing a segment across that position were counted and normalized by the total number of potential pairs. Within populations, the normalization factor was equal to $\binom{n}{2}$, where n is the population size. Between populations, it was the product of the respective sizes.

Sharing graphs. The amount of sharing for the analyzed dataset was visualized using the ShareViz software. Individuals were represented as nodes, grouped into populations of origin. The thickness of the edges between nodes represent the total amount of sharing (in

centimorgans) between each pair of individuals. For presenting populations geographically, planar quasi-isometric embedding (ISOMAP) was used, where distances between populations were defined as inverse of the populations' pair-wise average.

Sharing Statistics. To compute the average total sharing between populations I and J the following expression was used:

$W^{IJ} = \frac{\sum_{i \in I} \sum_{i \in J} w^{ij}}{nm}$ Where w^{ij} is the total sharing between individuals i and j from populations I and J respectively, n and m are the number of individuals in populations I and J. The average lengths of the shared segments across populations were computed through the arithmetic mean of the shared segments for each pair of populations. To compute the distribution of longest segments (Table 1, Fig. 3) the longest shared segments for all possible pairs was considered. The observed probability of a pair sharing a longest segment of a specified length was computed normalizing the observed counts by the number of possible pairs within or between the considered populations. The counts for all the histograms were obtained through floor rounding of the values.

Sharing between remote relatives. Siblings share, on average, the length of one haploid genome IBD. At each locus, sharing persists for an additional meiotic transmission with probability ½. Cousins therefore share a total of ¼ of the genome length on average, and k-th cousins share (¼)k of the genome. The length of a segment shared by k-th cousins is the length between adjacent crossover sites along any of the transmissions from the shared ancestor of the segment. This length is distributed exponentially, with mean inversely proportional to the number of transmissions involved. For k-th cousins, this mean is 50cM/(k+1).

Selecting loci with significantly excessive sharing. We defined a locus as excessively sharing if the frequency of shared segments there exceeded 4 standard deviations beyond the mean genomewide sharing.

Results

Jewish populations form distinctive clusters with genetic proximity to European and Middle Eastern groups. Affy v 6 data were generated for 237 unrelated individuals (51.1% female) from the 7 Jewish populations (table S1). To examine the population genetic structure of Jewish populations in the global and regional contexts, the SNP data were merged with selected datasets from the Human Genome Diversity Panel (HGDP). The first 2 principal components of worldwide populations showed that the Jewish populations clustered with the European groups (Fig. 1A). When compared only to the European and Middle Eastern, non-Jewish populations (Bedouins, Druze, Palestinians), each of the Jewish populations formed its own distinctive cluster, indicating the shared ancestry and relative genetic isolation of the members of each of those groups (Fig. 1B and 1C). Pairwise F_{ST} analysis indicated that each of these clusters was distinct and statistically different from all of the others (Tables 1 top, S2 and S3). ANOVA on the PCA Eigenvalues confirmed that the populations differed from one another ($p<0.0001$) as did the permutation testing of between-group IBD for all pair-wise comparisons of the 7 Jewish populations (tables S4 and S5). PC1 distinguished Northern and Southern European and Jewish and Middle Eastern populations. Along this axis, Europeans were closest to Ashkenazi Jews, followed by Sephardic, Italian, Syrian and Middle Eastern Jews. Of the European populations, the Northern Italians showed the greatest proximity to the Jews, followed by Sardinians and French (Fig. 1B), an observation that was confirmed by F_{ST} (Table 1). Also along this axis, the Adygei, a Caucasian population, showed proximity to the Ashkenazi Jews. The Druze, Bedouins

and Palestinians, respectively, were closest to the Middle Eastern (Iranian and Iraqi) and Syrian Jews (Fig. 1C). PC2 distinguished the Middle Eastern Jewish and non-Jewish populations (Fig. 1C). Along PC2, the clusters of the Iranian, Iraqi, and Syrian Jews and Druze, Bedouins and Palestinians followed a north to south distribution that was reminiscent of their geographic separation in the Middle East (Figs. 1B, 1C). Virtually identical results were observed when the Jewish groups were compared with the European national groups of the Population Reference Sample (PopRes) (fig. S1A and B). The observations with SNPs tended to be confirmed by CNVs. The principal component analysis of CNVs demonstrated distinctive clusters for all of the Jewish populations, except Iraqi Jews (Fig. 1D). The stability of these clusters was determined by using different numbers of CNVs, representing the tails of the genetic distance distributions (fig. S2).

These findings demonstrated that the most distant and differentiated of the Jewish populations were Iranian Jews followed by Iraqi Jews (average F_{ST} to all other Jewish populations 0.016 and 0.011, respectively). The closest genetic distance was between Greek and Turkish Sephardic Jews (F_{ST} = 0.001) who, in turn, were close to Italian, Syrian and Ashkenazi Jews. Thus, two major groups were identifiable that could be characterized as *Middle Eastern Jews* and *European/Syrian Jews*, an observation that was supported by pairwise F_{ST} and by phylogenetic tree analysis (Fig. 2C). Notably, the Iranian and Iraqi Jews were grouped together with strong statistical support. The European and Syrian Jews shared a common branch that included non-Jewish European populations. The Druze, Palestinian and Bedouins were on branches distinctive from the other populations. The robustness of this phylogenetic tree was demonstrated by the fact that a majority of major branching was supported by greater than 75% of bootstrap replications.

The Structure analysis was compatible with the Iranian and Iraqi Jews having predominant Middle Eastern/Central Asian ancestry and the European and Syrian Jews having both Middle Eastern/Central Asian and European ancestry with the proportion of European ancestry ranging between 20% and 40% when K ranged from 4 to 6. The Sephardic, Italian and Syrian Jews all showed a low level component (8-11%) that was shared with the North African Mozabite population when K equaleds 6 (Fig. 2A, 2B). This component was less apparent among the Ashkenazi and Middle Eastern Jews (Fig. 2B, fig. S3).

Jewish communities show high levels of IBD. IBD between Jewish individuals exhibited high frequencies of shared segments (Fig 3A, fig. S4). The median pair of individuals within a community shared a total of 50cM IBD (quartiles: 23.0cM and 92.6cM) -- such levels are expected to be shared by 4th or 5th cousins in a completely outbred population. However, the typical shared segments in these communities were shorter than expected between 5th cousins (8.33cM length), suggesting multiple lineages of more remote relatedness between most pairs of Jewish individuals (fig S5).

Within the different Jewish communities, three distinct patterns were observed (Fig. 3B, Table 1, fig. S4, S5). The Greek and Turkish Jews had relatively modest levels of IBD, similar to that observed in the French HGDP samples. The Italian, Syrian, Iranian and Iraqi Jews demonstrated the high levels of IBD that would be expected for extremely inbred populations. Unlike the other populations, the Ashkenazi Jews exhibited increased sharing of segments at the shorter end of the range (i.e. 5cM length), but decreased sharing at the longer end (i.e., 10cM). (fig. S5)

Frequent IBD between different Jewish populations reflects their genetic proximity. As expected, the vast majority of long shared segments (89% of 15cM segments, 78% of 10cM segments) were shared within communities. However, the genetic connections between the Jewish populations became evident from the frequent IBD across these Jewish groups (63% of all shared segments). The web of relatedness between the 27,966 pairs of individuals in this study was intricate, even if restricted only to the 2,166 pairs sharing a total 50cM or more, a level of sharing among third cousins (fig. S6). When population averages were examined, this network of IBD was consistent with the geographic distances between populations, with planar embedding (Fig. 3C) representing 93% of the initial information content. The notable exception was that of Turkish and Italian Jews who were nearest neighbors in terms of IBD, but more distant on the geographical map, potentially reflecting their shared Sephardic ancestry. Jewish populations shared more and longer segments with one another than with non-Jewish populations, highlighting the commonality of Jewish origin. Among pairs of populations ordered by total sharing, 12 out of the top 20 were pairs of Jewish populations, and none of the top 30 paired a Jewish population with a non-Jewish one (Fig. 3A).

Specific regions of the genome are frequently shared between Jewish populations. Shared regions spanned the entire genome, but none (longer than 5cM) was shared among all the Jewish populations. Between Jewish populations, spikes of frequently shared segments were observed relative to the lower background sharing (fig. S7 and S8). Loci that demonstrated significantly excessive ($\geq$4 standard deviations) sharing between Jewish populations are listed in table S6. These loci spanning >20 million bases in total, were not spanned by single LD blocks, nor did they include single haplotypes of high frequency (fig. S8). Gene content along these regions was slightly higher ($p<0.013$) than the genome wide average (table S6).

Timing of the Middle Eastern-European Jewish divergence. As a first step, population simulation was performed to estimate the ancestral population size for the Jewish and Middle-Eastern non-Jewish cohorts in this study. The ancient (before the introduction of agriculture, 500 years before present) ancestral population size was set to a smaller and realistic 1000 individuals per simulated population size, although this result does not change significantly, as the fraction of IBS pairs is affected almost exclusively by recent generations. Ashkenazi Jewish samples were excluded from this analysis, as the sharing in this population was inconsistent with a near-constant recent population size. The ancestral population size was then used in two simulations to estimate the time splits between Middle-Eastern and European (Italian) Jews. Under these assumptions, the split was consistent with 100-150 generations, or during the first millennium BCE, assuming a generation time of 20 years. The split between Middle-Eastern Jews and non-Jews was inconsistent with these simulation assumptions, suggesting a more complex history than a simple split of a single ancestral population.

Discussion

This study touches upon an issue that was raised over a century ago by Maurice Fishberg, Joseph Jacobs and others about whether the Jews constitute a race, a religious group or something else.[29-30] In this study, Jewish populations from the major Jewish Diaspora groups – Ashkenazi, Sephardic and Mizrahi – formed a distinctive population cluster by PCA analysis, albeit one that is closely related to European and Middle Eastern, non-Jewish populations. Within the study, each of the Jewish populations formed its own cluster as part of the larger Jewish cluster. Each group demonstrated Middle Eastern ancestry and variable admixture with European populations. This was observed in the Structure plots and in the *Fst* analysis by the proximity of all Jewish populations one to another, to non-Jewish Middle Eastern populations

and to non-Jewish Southern European (French, Northern Italian, and Sardinian) populations. The patterns of relatedness were similar, albeit with higher resolution to what was reported in a recent study of fewer Jewish populations using microsatellite markers.[31] Earlier investigators who studied fewer autosomal markers with less resolution and more recent investigators who studied Y chromosomal markers had similar observations. All noted that a major difference in Jewish groups was in the extent of admixture with local populations.[7-11,13,14,17]

Two major differences among the populations in this study were the high degree of European admixture (30-60%) among the Ashkenazi, Sephardic, Italian and Syrian Jews and the genetic proximity of these populations to each other compared to their proximity to Iranian and Iraqi Jews. This time of a split between Middle Eastern Iraqi and Iranian Jews and European/Syrian Jews, calculated by simulation and comparison of length distributions of IBD segments, is 100-150 generations, compatible with a historical divide that is reported to have occurred more than 2500 years ago.[2; 5] The Middle Eastern populations were formed by Jews in the Babylonian and Persian empires who are thought to have remained geographically continuous in those locales. In contrast, the other Jewish populations were formed more recently from Jews who migrated or were expelled from Palestine and from individuals who were converted to Judaism during Hellenic-Hasmonean times, when proselytism was a common Jewish practice. During Greco-Roman times, recorded mass conversions led to 6 million people practicing Judaism in Roman times or up to 10% of the population of the Roman Empire. Thus, the genetic proximity of these European/Syrian Jewish populations, including Ashkenazi Jews, to each other and to French, Northern Italian, and Sardinian populations favors the idea of non-Semitic Mediterranean ancestry in the formation of the European/Syrian Jewish groups and is incompatible with theories that Ashkenazi Jews are for the most part the direct lineal

descendants of converted Khazars or Slavs.[32] The genetic proximity of Ashkenazi Jews to southern European populations has been observed in several other recent studies.[33-36]

Admixture with local populations, including Khazars and Slavs, may have occurred subsequently during the 1000-year (2nd millennium) history of the European Jews. Based on analysis of Y chromosomal polymorphisms, Hammer estimated that the rate might have been as high as 0.5% per generation or 12.5% cumulatively (a figure derived from Motulsky), although this calculation might have underestimated the influx of European Y chromosomes during the initial formation of European Jewry.[15] Notably, up to 50% of Ashkenazi Jewish Y chromosomal haplogroups (E3b, G, J1 and Q) are of Middle Eastern origin,[15] whereas the other prevalent haplogroups (J2, R1a1 R1b) may be representative of the early European admixture.[20] The 7.5% prevalence of the R1a1 haplogroup among Ashkenazi Jews has been interpreted as a possible marker for Slavic or Khazar admixture because this haplogroup is very common among Ukrainians (where it was thought to have originated), Russians, and Sorbs, as well as among Central Asian populations, although the admixture may have occurred with Ukrainians, Poles or Russians, rather than Khazars.[12; 35] In support of the ancestry observations reported in the current study, the major distinguishing feature between Ashkenazi and Middle Eastern Jewish Y chromosomes was the absence of European haplogroups in Middle Eastern Jewish populations.[37] Four founder mitochondrial haplogroups of Middle Eastern origins comprise approximately 40% of the Ashkenazi Jewish genetic pool, whereas the remainder is comprised of other haplogroups, many of European origin and supporting the degree of admixture observed in the current study.[13] Evidence for founder females of Middle Eastern origin has been observed in other Jewish populations based on non-overlapping mitochondrial haplotypes with coalescence times >2000 years.[14] The number of founders and their relative proportions from one population to another is

variable. These Y chromosomal and mitochondrial haplogroup studies along with the population-specific genetic clusters and prevalent within and between-population IBD segments of the current study, and Mendelian genetic disease mutation studies all point to local founder effects with subsequent genetic drift that caused genetic differentiation.[38] The differential pattern of IBD observed only among Ashkenazi Jews in which older IBD segments became shorter and few new ones were created is consistent with a population bottleneck followed by rapid expansion (see Methods). This corresponds to the so-called *demographic miracle* of Ashkenazi Jewish history discussed earlier.[6]

The Iranian and Iraqi Jews are the most differentiated with the greatest genetic distances from the other populations and the least distances from each other, as well as the least sharing of the "European" component in *Structure*. Similar differentiation was observed for mitochondrial haplotypes.[14] The high rate of IBD within these groups (and in Italian and Syrian Jews) demonstrates a high coefficient of inbreeding. Yet, the sharing of Iranian and Iraqi Jews of a branch on the phylogenetic tree with the Adygei suggests that a certain degree of admixture may have occurred with local populations not included in this study.

Besides Southern European groups, the closest genetic neighbors to most Jewish populations are the Palestinians, Bedouins and Druze. The observed differentiation of these groups reflects their histories of within group endogamy.[39] Yet, their genetic proximity to one another and to European and Syrian Jews suggests a shared genetic history of related Middle Eastern and non-Semitic Mediterranean ancestors who chose different religious and tribal affiliations. These observations are supported by the significant overlap of Y chromosomal haplogroups between Israeli and Palestinian Arabs with Ashkenazi and non-Ashkenazi Jewish populations that has been described previously.[37] Likewise, a study comparing 20 microsatellite

markers in Israeli Jewish, Palestinians and Druze populations demonstrated the proximity of these 2 non-Jewish populations to Ashkenazi and Iraqi Jews.[40]

This study demonstrates that the studied Jewish populations represent a series of geographical isolates or clusters with genetic threads that weave them together. These threads are observed as IBD segments that are shared within and between Jewish groups. Over the past 3,000 years, both the flow of genes and the flow of religious and cultural ideas have contributed to Jewishness.

Acknowledgements

This study was supported, in part, by the Lewis and Rachel Rudin Foundation, the Iranian-American Jewish Federation, the U.S.-Israel Binational Science Foundation and private donors. The authors thank Alon Keinan, Alkes Price, Yongzhao Shao, Carlos Bustamante and David Goldstein for useful technical discussions; Lanchi U, Malka Sasson, Bridget Riley, Christopher Campbell, Jidong Shan, and David Reynolds for technical contributions; Lawrence Schiffman, Robert Chazen, Aron Rodrigue and Harvey Goldberg for informative historical discussions.

References

1. Biran, A., and Naveh, J. (1993). An Aramaic stele fragment from Tel Dan. Israel Exploration Journal 43, 81-98.
2. Ben-Sasson, H.H. (1976). A History of the Jewish People.(Cambridge: Harvard University Press).
3. Levy, H. (1999). Comprehensive History of the Jews of Iran.(Costa Mesa, CA: Mazda Publishers).
4. Rejwan, N. (1985). The Jews of Iraq: 3000 Years of History and Culture.(Boulder: Westview Press).
5. Baron, S.W. (1937). Social and Religious History of the Jews.(New York: Columbia University Press).
6. Weinryb, B. (1973). A History of the Jews in Poland.(Philadelphia: Jewish Publication Society of America).
7. Wexler, P. (2002). Two-Tiered Relexification in Yiddish: Jews, Sorbs, Khazars, and the Kiev-Polessian Dialect.(Berlin and New York: Mouton de Gruyter).
8. Cohen, S.J.D. (1999). The Beginnings of Jewishness.(Berkeley: University of California Press).
9. Livshits, G., Sokal, R.R., and Kobyliansky, E. (1991). Genetic affinities of Jewish populations. Am J Hum Genet 49, 131-146.
10. Karlin, S., Kenett, R., and Bonne-Tamir, B. (1979). Analysis of biochemical genetic data on Jewish populations: II. Results and interpretations of heterogeneity indices and distance measures with respect to standards. Am J Hum Genet 31, 341-365.
11. Carmelli, D., and Cavalli-Sforza, L.L. (1979). The genetic origin of the Jews: a multivariate approach. Hum Biol 51, 41-61.
12. Behar, D.M., Garrigan, D., Kaplan, M.E., Mobasher, Z., Rosengarten, D., Karafet, T.M., Quintana-Murci, L., Ostrer, H., Skorecki, K., and Hammer, M.F. (2004). Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum Genet 114, 354-365.
13. Behar, D.M., Metspalu, E., Kivisild, T., Achilli, A., Hadid, Y., Tzur, S., Pereira, L., Amorim, A., Quintana-Murci, L., Majamaa, K., et al. (2006). The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. Am J Hum Genet 78, 487-497.
14. Behar, D.M., Metspalu, E., Kivisild, T., Rosset, S., Tzur, S., Hadid, Y., Yudkovsky, G., Rosengarten, D., Pereira, L., Amorim, A., et al. (2008). Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora. PLoS One 3, e2062.
15. Hammer, M.F., Redd, A.J., Wood, E.T., Bonner, M.R., Jarjanazi, H., Karafet, T., Santachiara-Benerecetti, S., Oppenheim, A., Jobling, M.A., Jenkins, T., et al. (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, 6769-6774.
16. Nebel, A., Filon, D., Faerman, M., Soodyall, H., and Oppenheim, A. (2005). Y chromosome evidence for a founder effect in Ashkenazi Jews. Eur J Hum Genet 13, 388-391.
17. Oddoux, C., Guillen-Navarro, E., Ditivoli, C., Dicave, E., Cilio, M.R., Clayton, C.M., Nelson, H., Sarafoglou, K., McCain, N., Peretz, H., et al. (1999). Mendelian diseases among Roman Jews: implications for the origins of disease alleles. J Clin Endocrinol Metab 84, 4405-4409.

18. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science 298, 2381-2385.
19. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., et al. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83, 347-358.
20. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100-1104.
21. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet 85, 762-774.
22. Weir, B.S., and Cockerham, C.C. (1984). Evolution 38, 1358-1370.
23. Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-1599.
24. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet 2, e190.
25. Jakobsson, M., and Rosenberg, N.A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23, 1801-1806.
26. Rosenberg, N.A. (2004). Distruct: a program for the graphical display of population structure. Molecular Ecology Notes 4, 137-138.
27. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res 19, 318-326.
28. Liang, L., Zollner, S., and Abecasis, G.R. (2007). GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics 23, 1565-1567.
29. Fishberg, M. (1911). The Jews: A Study of Race and Environment.(New York: Charles Scribner's Sons).
30. Jacobs, J. (1891). Jewish Statistics: Social, Vital, Anthropometric.(London: D. Nutt).
31. Kopelman, N.M., Stone, L., Wang, C., Gefel, D., Feldman, M.W., Hillel, J., and Rosenberg, N.A. (2009). Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. BMC Genet 10, 80.
32. Koestler, A. (1976). The Thirteenth Tribe.(New York: Random House).
33. Tian, C., Plenge, R.M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A.E., Qi, L., Gregersen, P.K., et al. (2008). Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 4, e4.
34. Price, A.L., Butler, J., Patterson, N., Capelli, C., Pascali, V.L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A.A., Korkolopoulou, P., et al. (2008). Discerning the ancestry of European Americans in genetic association studies. PLoS Genet 4, e236.
35. Need, A.C., Kasperaviciute, D., Cirulli, E.T., and Goldstein, D.B. (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, R7.

36. Olshen, A.B., Gold, B., Lohmueller, K.E., Struewing, J.P., Satagopan, J., Stefanov, S.A., Eskin, E., Kirchhoff, T., Lautenberger, J.A., Klein, R.J., et al. (2008). Analysis of genetic variation in Ashkenazi Jews by high density SNP genotyping. BMC Genet 9, 14.
37. Nebel, A., Filon, D., Brinkmann, B., Majumder, P.P., Faerman, M., and Oppenheim, A. (2001). The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. Am J Hum Genet 69, 1095-1112.
38. Ostrer, H. (2001). A genetic profile of contemporary Jewish populations. Nat Rev Genet 2, 891-898.
39. Curtis, M., and American Academic Association for Peace in the Middle East. (1975). The Palestinians : people, history, politics.(New Brunswick, N.J.: Transaction Books).
40. Rosenberg, N.A., Woolf, E., Pritchard, J.K., Schaap, T., Gefel, D., Shpirer, I., Lavi, U., Bonne-Tamir, B., Hillel, J., and Feldman, M.W. (2001). Distinctive genetic signatures in the Libyan Jews. Proc Natl Acad Sci U S A 98, 858-863.
41. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575.
42. Felsenstein, J. (1989) PHYLIP -- Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166

Figure Legends

Figure 1. Principal components analysis of Jewish populations in a global (A) and regional context (B, PC1 vs. PC2; C, PC1 vs. PC3) CNVs (D, PC1 vs. PC3). ASH, Ashkenazi Jews; IRN, Iranian Jews; IRQ, Iraqi Jews; SYR, Syrian Jews; ITJ, Italian Jews; GRK, Sephardic Greek Jews; TUR, Sephardic Turkey Jews. N. Italian is a combined set comprising Bergamo and Tuscan Italians. In A, Middle Eastern non-Jewish populations are in blue, Jewish populations are in brown and European populations are in red.

Figure 2. STRUCTURE and phylogenetic analysis of Jewish populations. (A) STRUCTURE results for K=2 to 6 for Jewish populations combined with selected HGDP worldwide populations. Each individual is represented by a vertical line, partitioned into colored segments that correspond to membership coefficients in the subgroups. The analysis is based on 3,904 SNPs with potentially high informativeness in revealing population structure (see Methods). (B) Expanded view of STRUCTURE results for Jewish populations for K=4 to 6. (C) Neighbor-joining tree of Jewish, European, and Israel non-Jews populations with Central/Southern African population as outgroup. Pairwise *Fst* distances were used for constructing the tree. Major population groups are indicated by right bracket. 500 bootstrap replications were performed to obtain confidence value for each interior node. Only bootstrap values above 50% are shown.

Figure 3. A) Average total sharing across populations. The genome-wide average IBD sharing (Y axis) for any two individuals sampled from different Mediterranean and European population pairs (X axis: top 50% sharing pairs, detail on top 15% pairs) was computed. The population pairs have been grouped into Jewish-Jewish (red bars), Jewish-non Jewish (yellow

bars) and non Jewish-non Jewish (blue bars). B) Distribution of segment lengths within each Jewish population. The expected number of IBD segments shared within each Jewish population (Y axis) for segments of length 5 cM and 10 cM were computed. C) Planar embedding of Jewish populations, with their inverse distances corresponding to average IBD between them (see Methods).

Table 1

Table 1. Genetic diversity of Jewish, European, and Middle Eastern non-Jewish populations. Pairwise *Fst*[b] is shown in the upper triangle. Pairwise sharing distance between populations, defined as the total centiMorgan length of IBD segments >3cM each averaged across all pairs of samples from the respective populations, is shown in the lower triangle.

Populations	*N*	H_o[a]	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR	N. Italian	Sardinian	French	Basque	Adygei	Russian	Palestinian	Druze	Bedouin
IRN	**28**	0.291		**0.015**	**0.015**	**0.017**	**0.018**	**0.015**	**0.014**	**0.018**	**0.027**	**0.022**	**0.030**	**0.018**	**0.028**	**0.017**	**0.017**	**0.021**
IRQ	**37**	0.293	**4.906**		**0.008**	**0.013**	**0.012**	**0.009**	**0.008**	**0.012**	**0.019**	**0.016**	**0.024**	**0.013**	**0.023**	**0.010**	**0.012**	**0.015**
SYR	**25**	0.296	**0.999**	**3.145**		**0.008**	**0.008**	**0.004**	**0.003**	**0.007**	**0.014**	**0.010**	**0.018**	**0.010**	**0.018**	**0.007**	**0.009**	**0.012**
ASH	**34**	0.294	**0.746**	**0.827**	**1.926**		**0.009**	**0.006**	**0.005**	**0.008**	**0.014**	**0.009**	**0.017**	**0.012**	**0.016**	**0.011**	**0.012**	**0.016**
ITJ	**37**	0.294	**0.609**	**0.857**	**1.566**	**3.093**		**0.005**	**0.005**	**0.008**	**0.014**	**0.011**	**0.018**	**0.012**	**0.018**	**0.010**	**0.011**	**0.015**
GRK	**42**	0.296	**0.564**	**0.773**	**1.570**	**2.153**	**2.476**		**0.001**	**0.004**	**0.010**	**0.007**	**0.014**	**0.009**	**0.015**	**0.006**	**0.008**	**0.011**
TUR	**34**	0.297	**0.747**	**1.043**	**2.049**	**2.954**	**2.411**	**2.556**		**0.004**	**0.010**	**0.007**	**0.014**	**0.008**	**0.014**	**0.005**	**0.007**	**0.010**
N_Italian	**21**	0.295	**0.675**	**0.740**	**0.865**	**1.015**	**0.978**	**0.906**	**0.899**		**0.007**	**0.002**	**0.008**	**0.008**	**0.009**	**0.010**	**0.011**	**0.016**
Sardinian	**28**	0.289	**0.675**	**0.683**	**0.970**	**1.098**	**0.852**	**0.955**	**0.946**	**1.386**		**0.009**	**0.013**	**0.019**	**0.020**	**0.017**	**0.017**	**0.022**
French	**28**	0.296	**0.498**	**0.623**	**0.999**	**1.012**	**0.948**	**0.889**	**0.937**	**1.361**	**1.353**		**0.007**	**0.009**	**0.005**	**0.014**	**0.014**	**0.020**
Basque	**24**	0.291	**0.584**	**0.662**	**0.854**	**1.153**	**0.862**	**0.935**	**0.905**	**1.427**	**1.472**	**2.078**		**0.018**	**0.015**	**0.021**	**0.021**	**0.027**
Adygei	**17**	0.298	**0.604**	**0.504**	**0.655**	**0.738**	**0.748**	**0.805**	**0.699**	**0.840**	**0.647**	**0.831**	**1.073**		**0.012**	**0.012**	**0.012**	**0.019**
Russian	**25**	0.295	**0.470**	**0.524**	**0.623**	**0.913**	**0.822**	**0.642**	**0.811**	**1.236**	**0.933**	**1.460**	**1.205**	**0.905**		**0.021**	**0.021**	**0.028**
Palestinian	**39**	0.303	**0.530**	**0.642**	**0.597**	**0.580**	**0.659**	**0.609**	**0.708**	**0.514**	**0.670**	**0.549**	**0.590**	**0.562**	**0.480**		**0.009**	**0.009**
Druze	**36**	0.296	**0.656**	**0.638**	**0.754**	**0.778**	**0.671**	**0.738**	**0.752**	**0.658**	**0.713**	**0.595**	**0.675**	**0.797**	**0.568**	**0.623**		**0.013**
Bedouin	**40**	0.301	**0.572**	**0.606**	**0.564**	**0.576**	**0.567**	**0.541**	**0.590**	**0.606**	**0.545**	**0.529**	**0.542**	**0.386**	**0.371**	**1.013**	**0.649**	
Total Sharing			**41.947**	**33.360**	**17.261**	**11.620**	**28.446**	**6.005**	**4.458**	**2.366**	**10.839**	**1.629**	**15.966**	**6.285**	**5.799**	**25.504**	**49.590**	**25.361**

[a] H_o is observed heterozygosity

[b]Confidence interval are listed in table S2

Figure 1

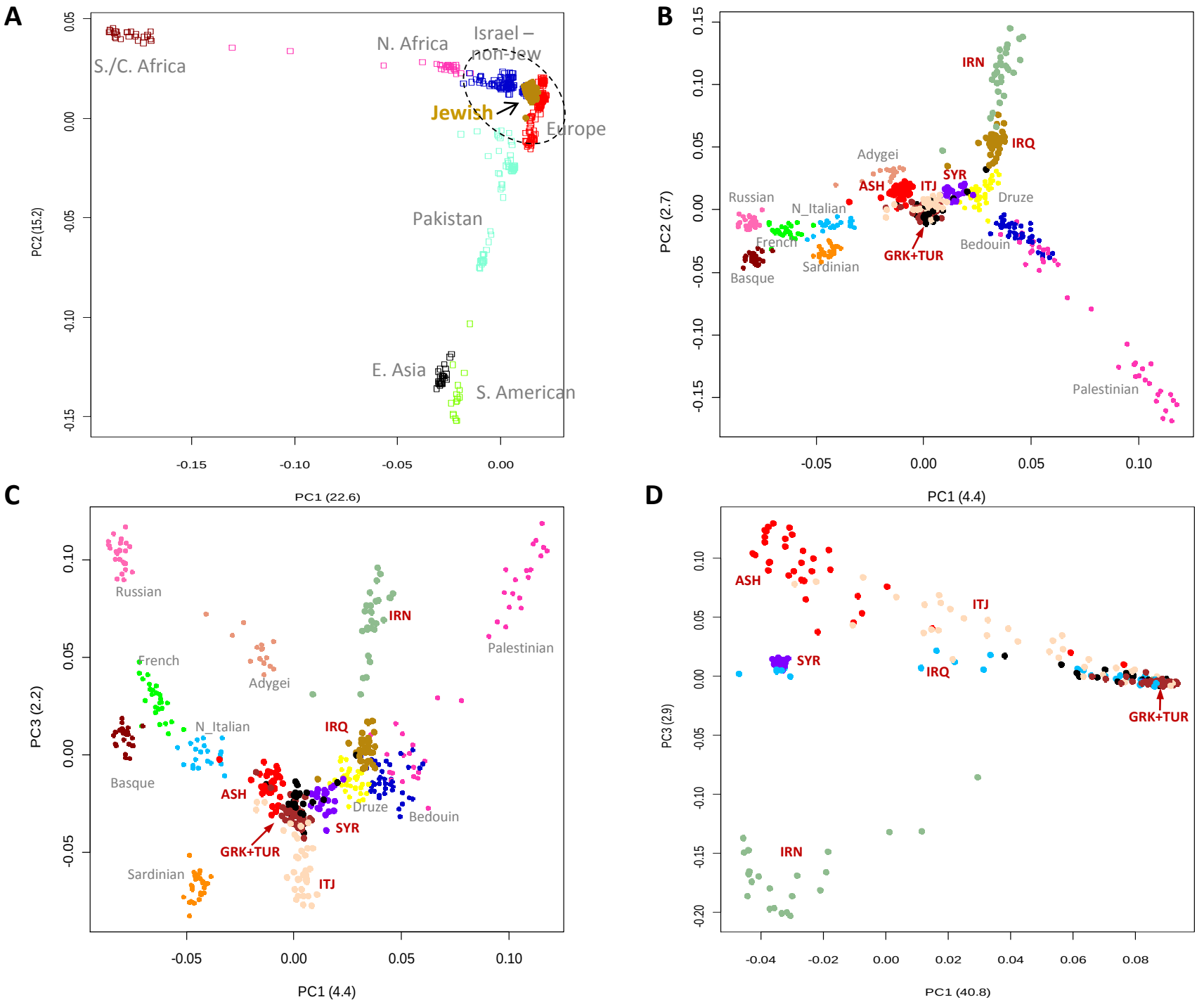

A
S./C. Africa
N. Africa
Israel –
non-Jew
Jewish
Europe
Pakistan
E. Asia
S. American
PC2 (15.2)
PC1 (22.6)
B
IRN
IRQ
Adygei
SYR
ASH
ITJ
Russian
N_Italian
Druze
French
Bedouin
GRK+TUR
Sardinian
Basque
Palestinian
PC2 (2.7)
PC1 (4.4)
C
Russian
IRN
Palestinian
French
Adygei
N_Italian
IRQ
Basque
ASH
Druze
Bedouin
SYR
GRK+TUR
Sardinian
ITJ
PC3 (2.2)
PC1 (4.4)
D
ASH
ITJ
SYR
IRQ
GRK+TUR
IRN
PC3 (2.9)
PC1 (40.8)

Figure 2
Figure 2

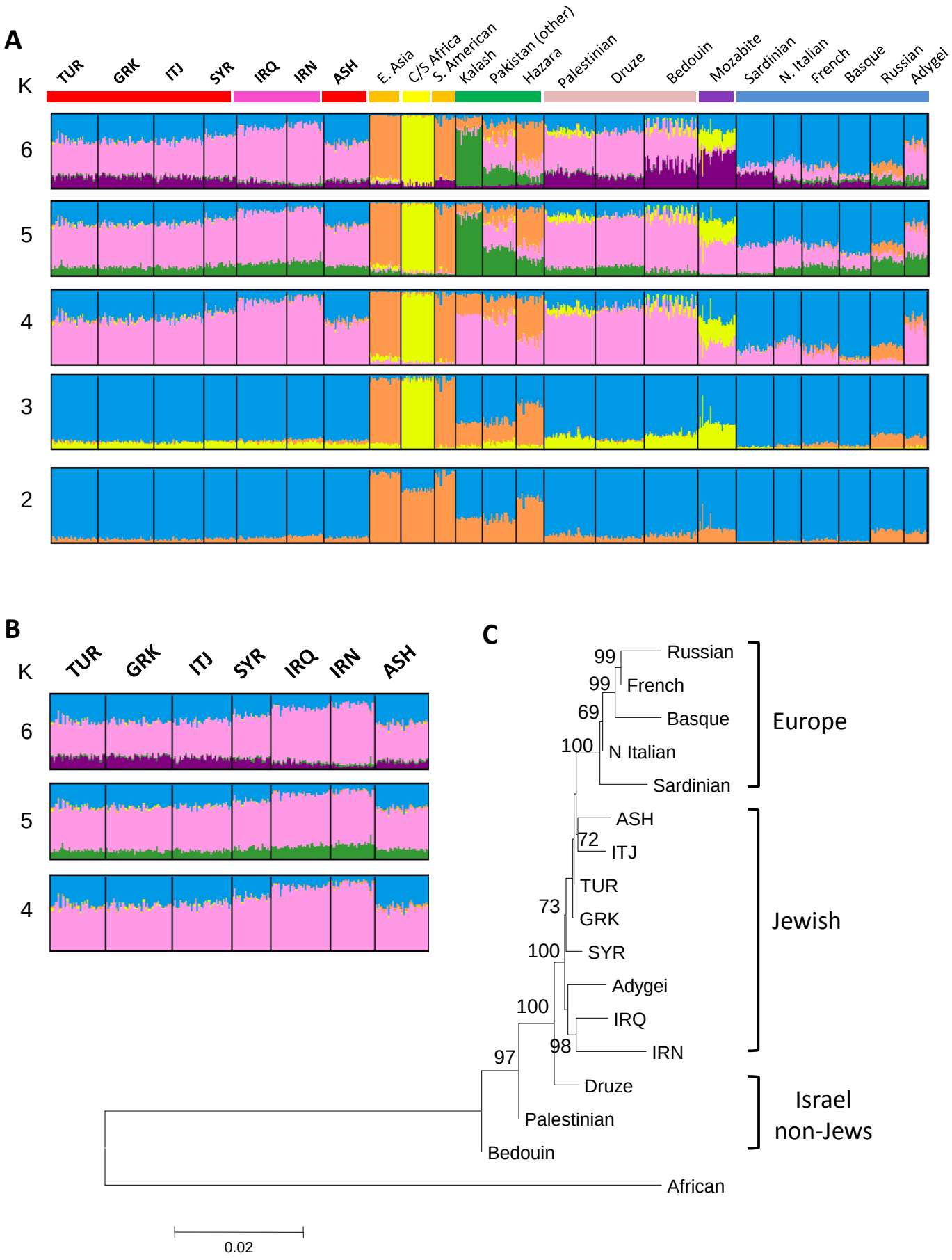
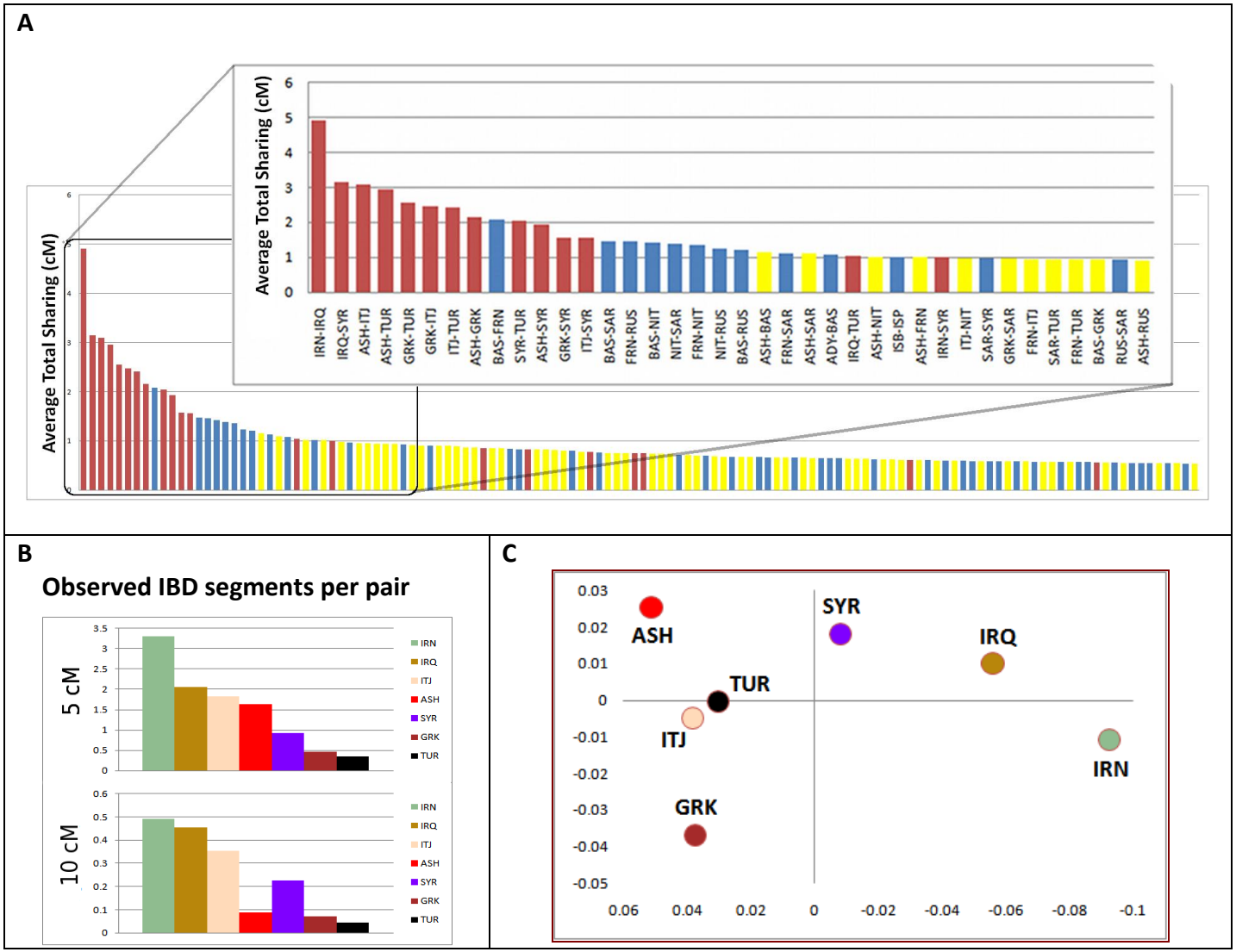


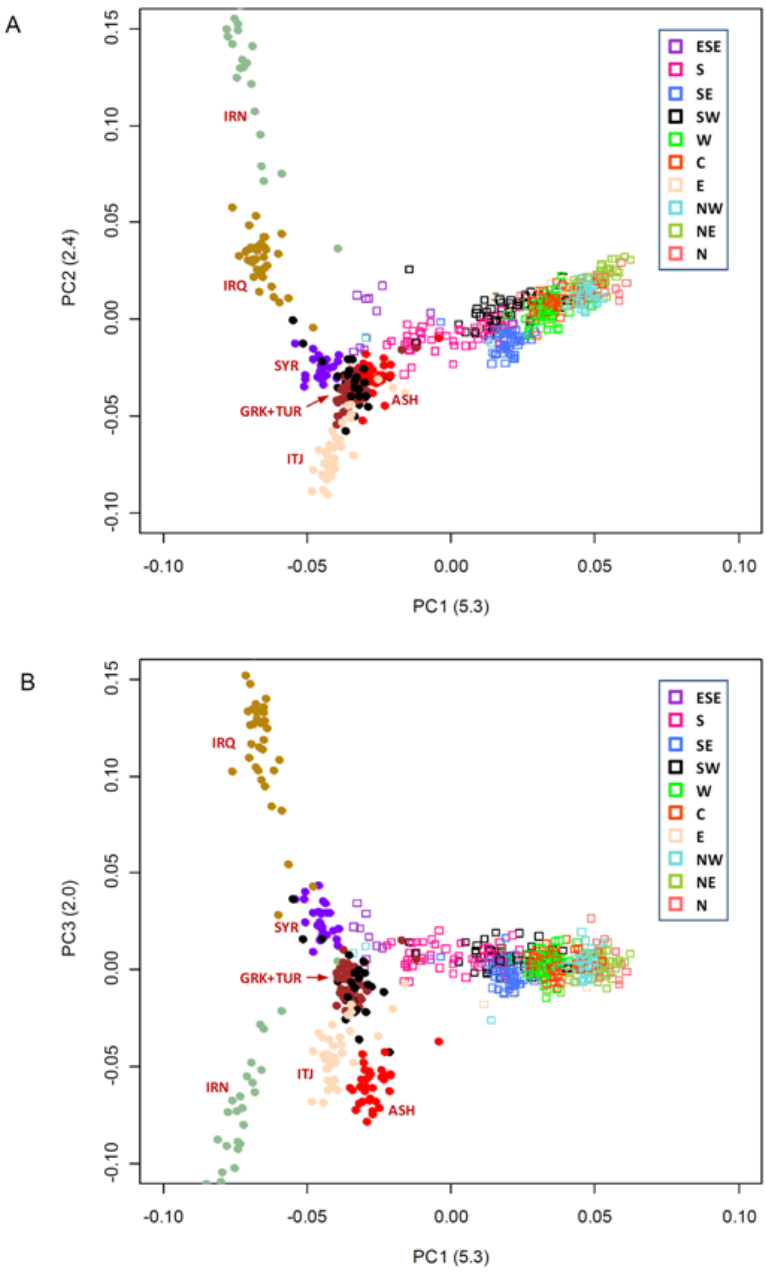
Figure 3

Figure 3.

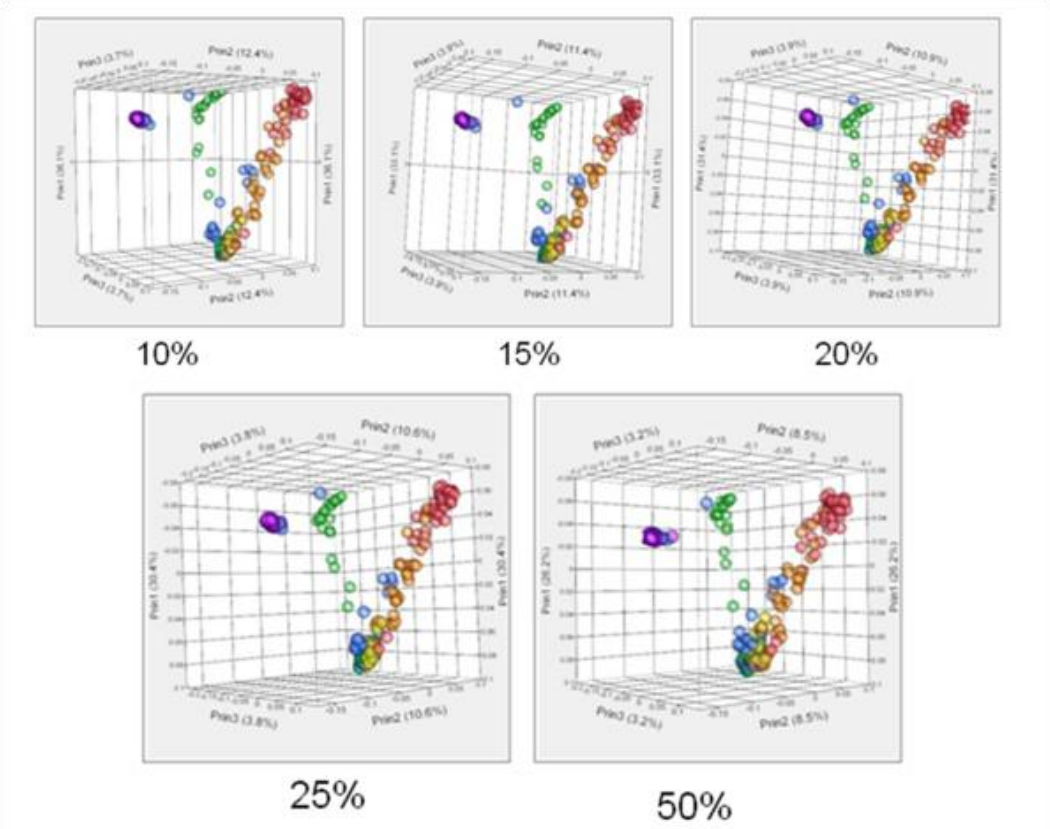


Supplemental figures

Supplemental figure 1. Principal component analysis of Jewish populations combined with selected PopRes populations. The Jewish samples are represented by large filled circles. Colored open squares represent PopRes samples grouped by different geographical origin in Europe. ESE, East Southeastern; S, Southern; SE, Southeastern; SW, Southwestern; W, Western; C, Central; E, Eastern; NW, Northwestern; NE, Northeastern; N, Northern. The list of countries from each geographical group can be found in supplemental methods.

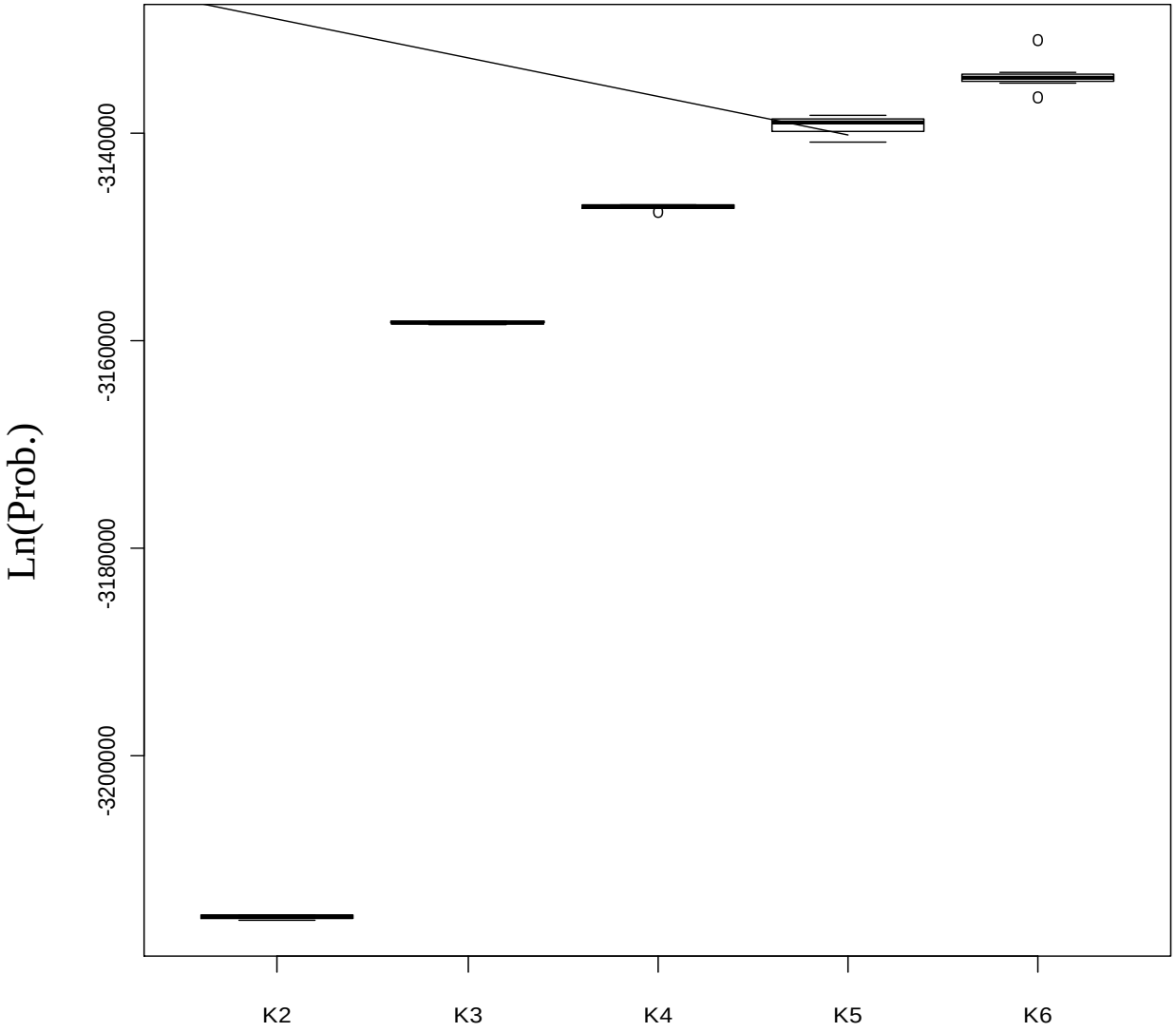


Supplement figure 2. PCA calculation within selected tails (i.e 10,15,20,25 and 50% of genetic distance distribution) of total genomic CNVs (946k).

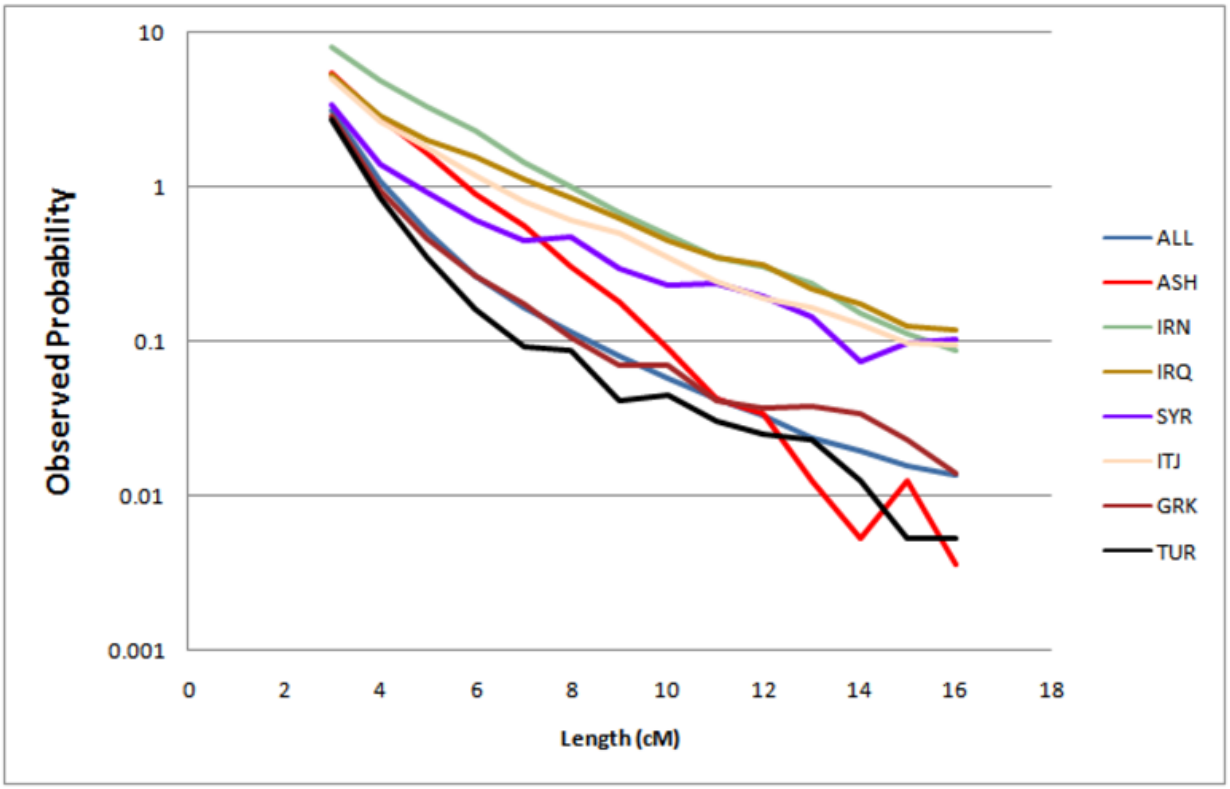


		Percentage of tail				
Comparison		10%	15%	20%	25%	50%
ASH	ITG	0.051755	0.045278	0.039793	0.03607	0.014566
ASH	IRN	0.052773	0.05235	0.049922	0.04612	0.021964
ASH	TUR	0.063995	0.054571	0.047573	0.042891	0.016607
ASH	GRK	0.064175	0.054724	0.047732	0.043039	0.016551
ASH	IRQ	0.073835	0.062447	0.054152	0.048615	0.018571
ASH	SYR	0.108403	0.091458	0.079234	0.070847	0.025553
IRN	ITG	0.046877	0.046482	0.044022	0.040202	0.01761
IRN	GRK	0.055216	0.052455	0.048884	0.04436	0.018537
IRN	TUR	0.056173	0.053436	0.049808	0.045205	0.018948
IRN	IRQ	0.065271	0.06072	0.055824	0.05044	0.020805
IRN	SYR	0.099149	0.089314	0.080597	0.072381	0.027672
IRQ	GRK	0.019054	0.015514	0.013151	0.011619	0.004213
IRQ	TUR	0.019208	0.015668	0.013296	0.011739	0.004432
IRQ	ITG	0.026824	0.022336	0.018964	0.016766	0.00588
IRQ	SYR	0.041908	0.034677	0.029802	0.026364	0.009057
ITG	GRK	0.014398	0.012186	0.010378	0.009195	0.003013
ITG	TUR	0.014682	0.012544	0.010737	0.009539	0.003294
ITG	SYR	0.061458	0.051426	0.044088	0.039007	0.012744
GRK	TUR	0.002753	0.00244	0.002193	0.002	0.000879
GRK	SYR	0.04946	0.041349	0.035589	0.031503	0.010404
SYR	TUR	0.049693	0.041624	0.03586	0.031769	0.010632

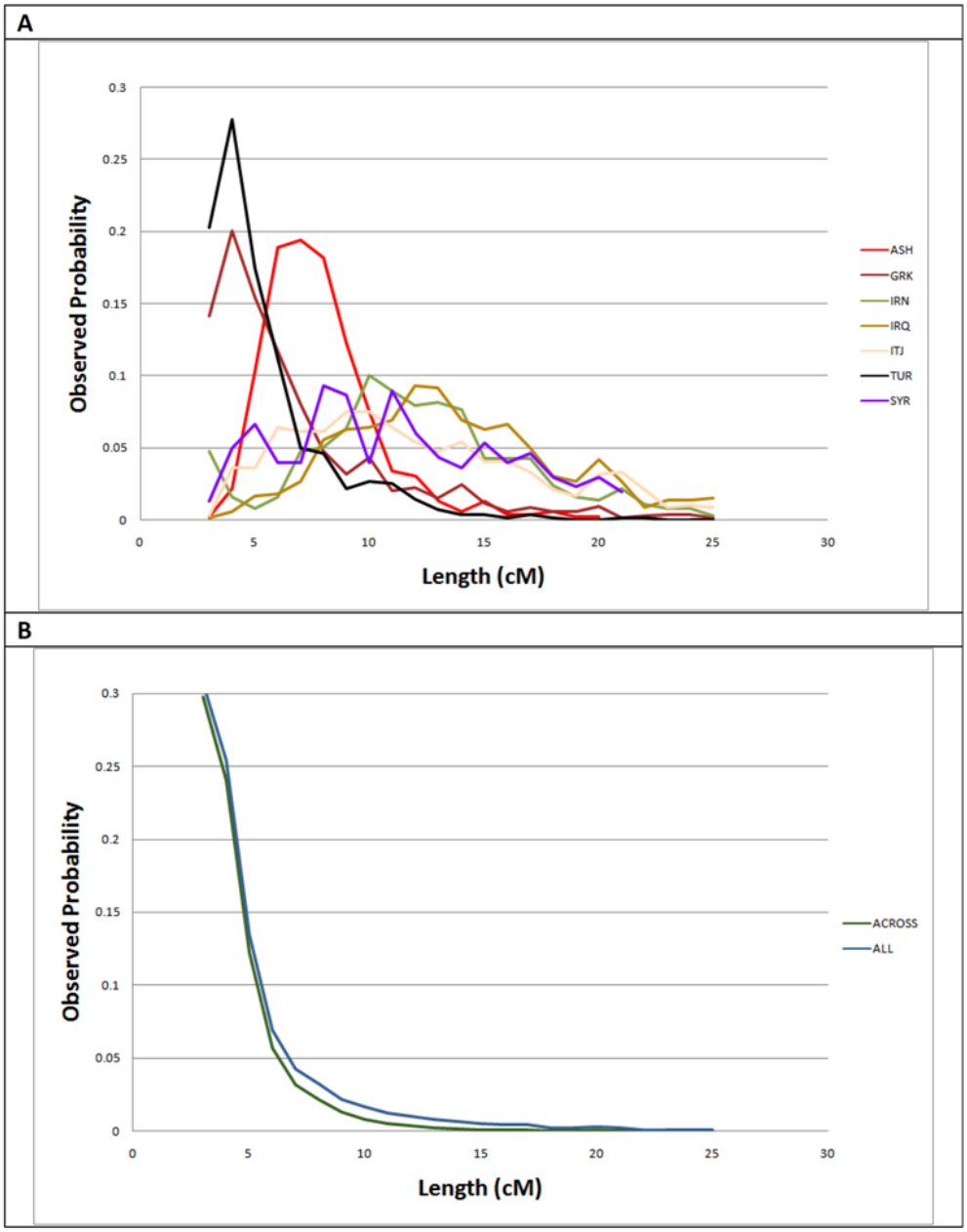
Supplemental figure 3. Likelihood scores for 10 iterations for Jewish and selected HGDP samples for K=2-6.



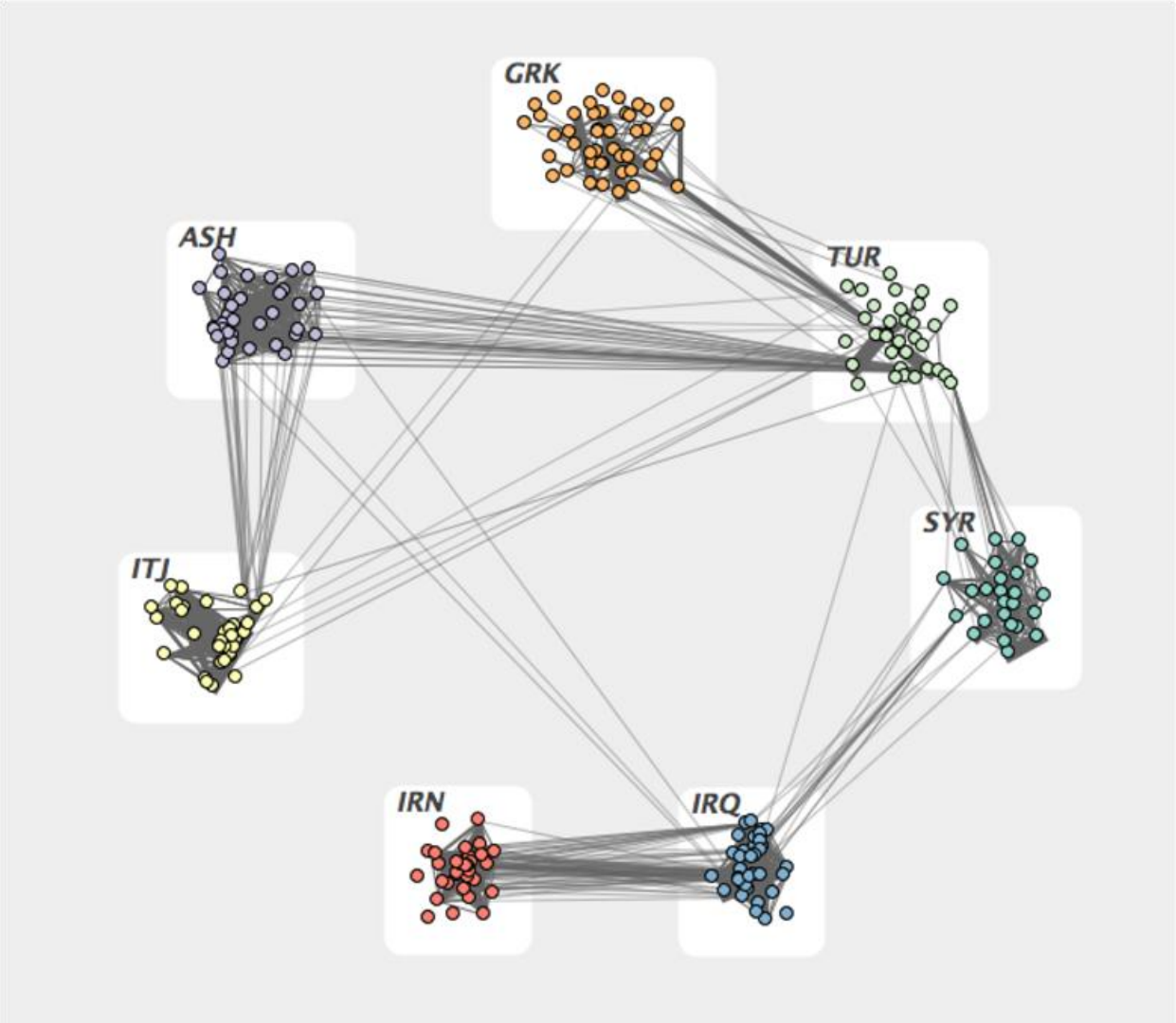
Supplemental figure 4. Distribution of segment lengths within each Jewish population. The expected number of IBD segments shared within each Jewish population (Y axis, logarithmic scale) for the discrete segment length range of 3 cM to 16 cM (X axis) were computed. An exponential decay rate (reflected by a linear behavior in the logarithmic scale) is representative of a recent population bottleneck. The behavior exhibited by the Ashkenazi population is consistent with the historical reports of a severe bottleneck followed by a rapid expansion, whereas the Greek and Turkish population decay suggests a more outbred profile.



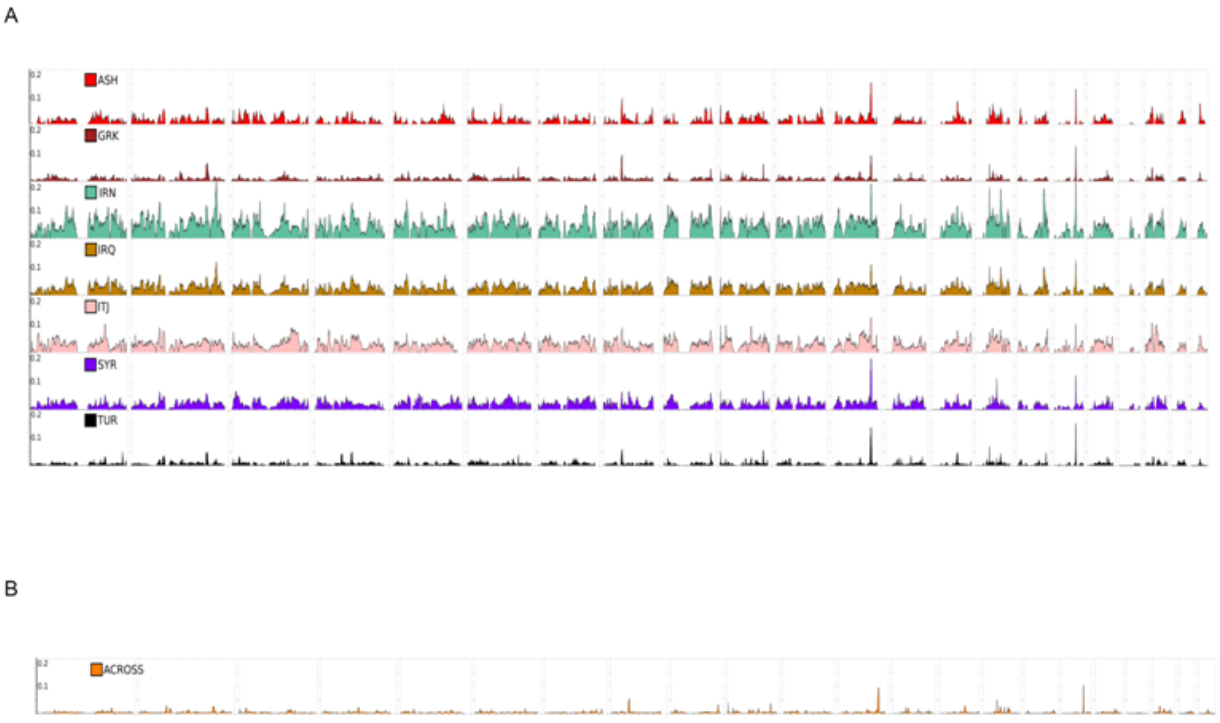
Supplemental figure 5. Distribution of longest IBD segments. The longest segments shared by the average pair of individuals as a function of the segment's length were computed. This distribution provides insights on the population's average time to most recent common ancestor. (A) Distribution of longest segments within each Jewish population. The average time to most common recent ancestor is larger for Turkish and Greek population, suggesting a more outbred profile than the other Jewish groups. (B) Distribution of longest segments for all possible pairs of Jewish individuals compared to the distribution of longest segments limited to pairs from different populations. The distribution for all possible pairs of individuals is shifted towards longer segments, reflecting the presence of closer relationships within populations. Short length for longest segments is expected across populations.



Supplemental figure 6. Network of relatedness between individuals. Each individual is represented by a node. Nodes are linked if the corresponding individuals are sharing >50cM of their genomes IBD.

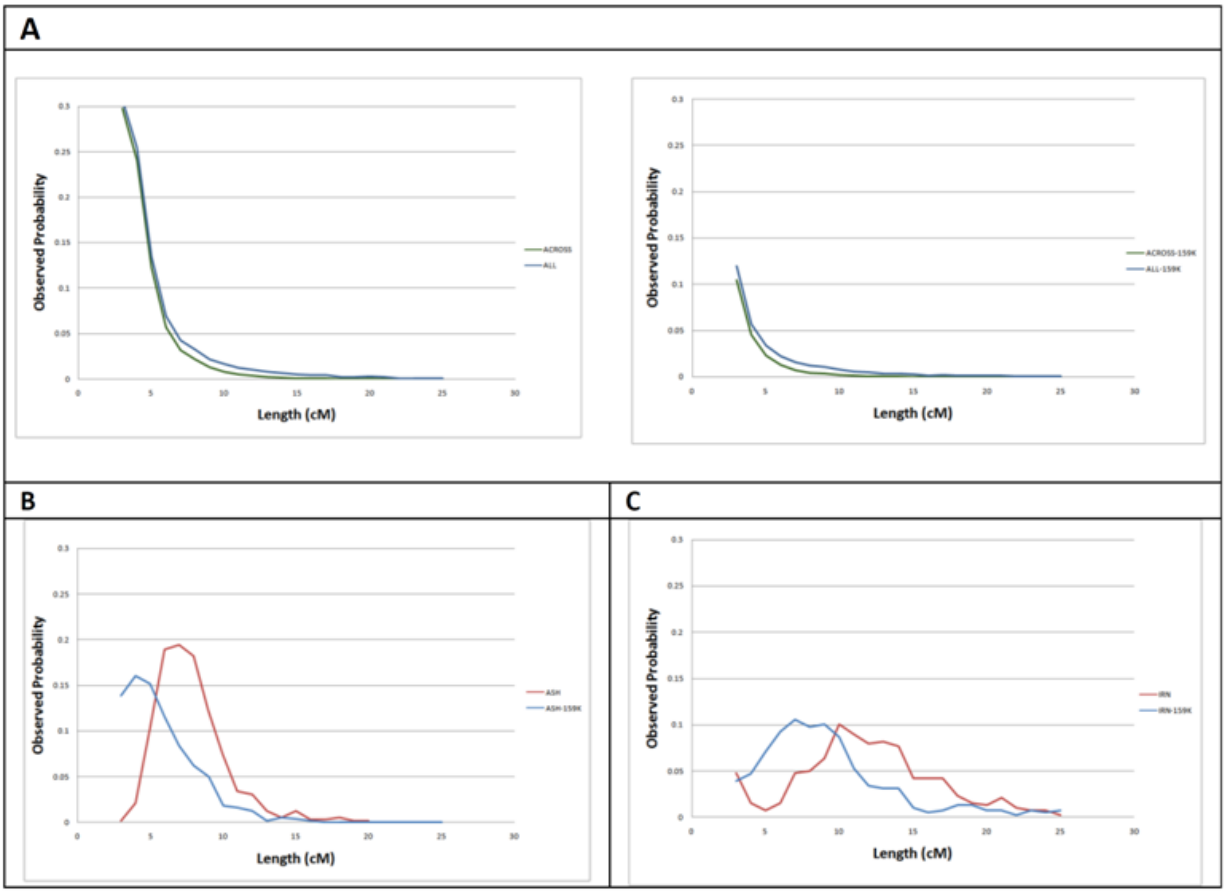


Supplemental figure 7. Manhattan plots of IBD sharing in Jewish populations. Normalized amount of sharing (Y axis) shown as a function of genomic position (X axis). (A) Sharing within populations for each Jewish group. (B) Sharing across populations.

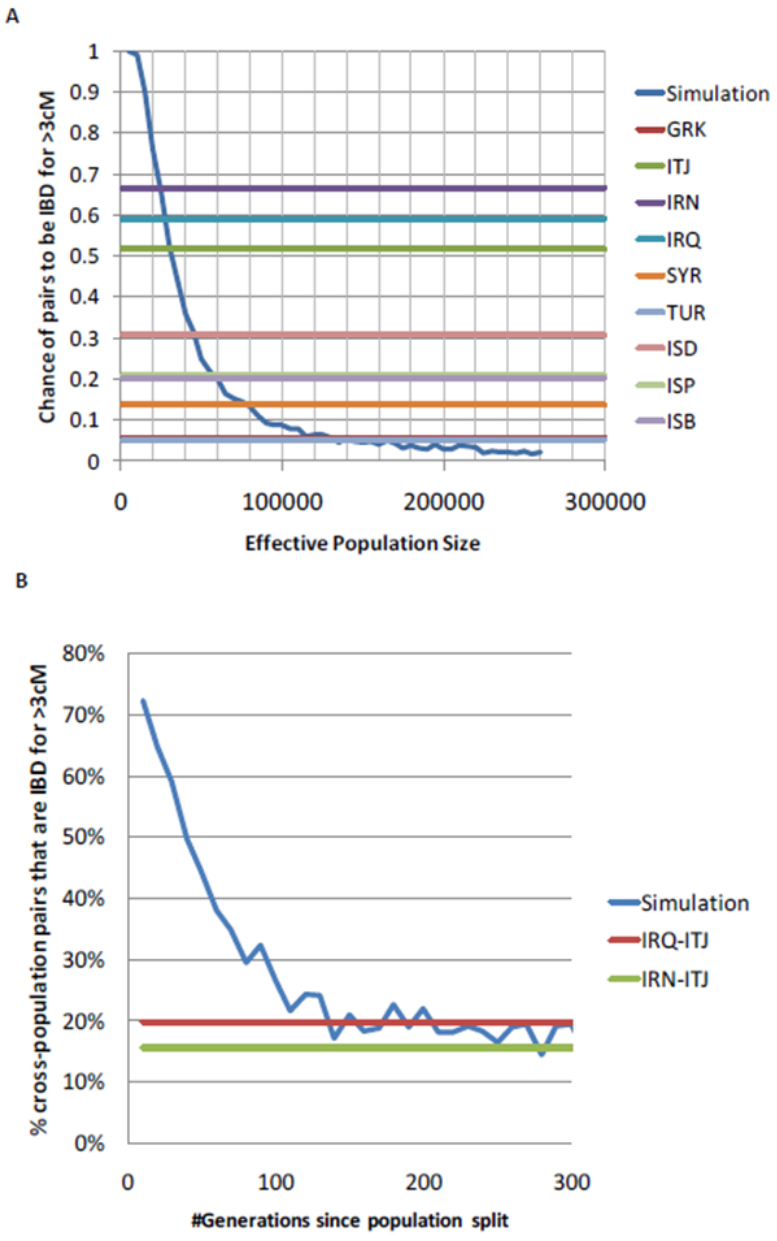


Supplemental figure 8. LD structure of IBD sharing peaks. Regions of intense sharing across Jewish populations were processed using the Haploview software package. The structure of LD blocks in the analyzed regions does not justify the observed levels of IBD sharing.

Supplemental figure 9 – Evaluation of power for IBD discovery based on a sparse set of SNPs. The results of the IBD analysis for the Jewish populations using the full set of markers and the sparser set of SNPs from the platforms' consensus were compared. (A) Distribution of longest segments for all possible Jewish pairs and limited to pairs across different populations. Comparison of the results for the full set (left) and the sparse set of markers (right). (B-C) Comparison of the results for the distribution of longest segments in the Ashkenazi and Iranian populations. Qualitatively similar results are obtained. Stricter parameters were employed in the IBD discovery phase for the sparser set of markers: long segments were excised, shifting the distributions towards shorter segments.



Supplemental figure 10. Inference of population history based on the observed fraction of individual pairs sharing a segment IBD (Y-axis) (A) First stage: simulation of a single population changing the ancestral population size (X-axis) to estimate the size appropriate for the Jewish and Middle Eastern non-Jewish populations. (B) Second stage : 2-population simulations changing the time of splits between Middle Eastern Jews and European (Italian) Jews is consistent with a population split 100-150 generations ago; the split between either of these groups and non-Jews is inconsistent with these simulation assumptions, as it predicts a distribution of shared segment lengths that does not fit any such scenario (data not shown), suggesting a more complex history than a simple split if a single ancestral population.[38]



Supplemental table 1. - Individuals genotyped, including country of origin and gender.

JHM ID	Pop.	Gender	JHM ID	Pop.	Gender	JHM ID	Pop.	Gender
3	IRQ	Female	307	SYR	Male	626	GRK	Female
4	IRQ	Male	308	SYR	Female	627	GRK	Male
5	IRQ	Male	311	SYR	Female	628	GRK	Male
6	IRQ	Male	312	SYR	Male	631	GRK	Female
7	IRQ	Female	313	SYR	Male	632	GRK	Female
15	IRQ	Female	315	SYR	Male	633	GRK	Male
17	IRQ	Female	316	SYR	Female	634	GRK	Female
21	IRQ	Female	317	SYR	Female	635	GRK	Male
23	IRQ	Female	318	SYR	Female	636	GRK	Female
24	IRQ	Male	323	SYR	Male	637	GRK	Female
28	IRQ	Female	325	SYR	Female	638	GRK	Male
29	IRQ	Female	330	SYR	Female	640	GRK	Female
30	IRQ	Male	331	SYR	Male	642	GRK	Male
32	IRQ	Male	332	SYR	Female	643	GRK	Female
33	IRQ	Male	333	SYR	Male	644	GRK	Female
36	IRQ	Male	336	SYR	Male	646	GRK	Male
37	IRQ	Female	341	SYR	Male	647	GRK	Male
40	IRQ	Female	351	SYR	Male	650	GRK	Male
52	TUR	Male	355	SYR	Female	651	GRK	Female
58	TUR	Male	358	SYR	Male	652	GRK	Male
59	TUR	Male	367	SYR	Male	654	GRK	Female
66	TUR	Female	368	SYR	Female	655	GRK	Male
68	TUR	Male	370	SYR	Female	657	GRK	Female
70	TUR	Female	371	IRQ	Female	662	GRK	Female
71	TUR	Male	375	IRQ	Female	663	GRK	Male
72	TUR	Male	378	SYR	Male	665	GRK	Female
76	TUR	Female	393	IRQ	Male	901	ITJ	Male
82	IRN	Female	399	IRQ	Male	902	ITJ	Female
83	IRN	Female	400	IRQ	Female	904	ITJ	Male
85	IRN	Male	402	IRQ	Male	909	ITJ	Female
87	IRN	Female	403	IRQ	Female	912	ITJ	Male
101	IRN	Male	440	IRQ	Female	913	ITJ	Female
104	IRN	Male	455	IRQ	Female	916	ITJ	Male
106	IRN	Female	457	IRQ	Male	917	ITJ	Female
129	IRN	Male	600	GRK	Male	1051	TUR	Female
136	IRN	Female	601	GRK	Female	1052	TUR	Female
148	IRN	Male	604	GRK	Male	1053	TUR	Female
151	IRN	Female	605	GRK	Male	1056	TUR	Female
157	IRN	Male	606	GRK	Male	1057	TUR	Female
158	IRN	Female	607	GRK	Female	1058	TUR	Female
159	IRN	Female	609	GRK	Male	1059	TUR	Female
164	IRN	Female	610	GRK	Male	1060	TUR	Male
167	IRN	Male	611	GRK	Male	1061	TUR	Male
180	IRN	Female	612	GRK	Female	1062	TUR	Female
188	IRN	Male	613	GRK	Male	1064	TUR	Female
196	IRN	Male	616	GRK	Male	1065	TUR	Female
268	IRN	Male	619	GRK	Male	1066	TUR	Female
273	IRN	Male	622	GRK	Male	1067	TUR	Female
276	IRN	Female	623	GRK	Female	1068	TUR	Male
305	SYR	Female	624	GRK	Female	1073	TUR	Female

Supplement Table 2. 95% confidence intervals for pairwise *Fst* between individual Jewish populations and between individual Jewish populations and populations in the HGDP panel

	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR	N_Italian	Sardinian	French	Basque	Adygei	Russian	Palestinian	Druze	Bedouin
IRN		0.01577-0.01589	0.01618-0.01632	0.01636-0.01648	0.01886-0.01901	0.01518-0.01530	0.01521-0.01534	0.01925-0.01940	0.02839-0.02858	0.02229-0.02244	0.03074-0.03094	0.01763-0.01779	0.02835-0.02853	0.01633-0.01645	0.01700-0.01712	0.02163-0.02176
IRQ			0.00801-0.00811	0.01358-0.01370	0.01215-0.01225	0.00849-0.00858	0.00827-0.00836	0.01202-0.01215	0.02053-0.02067	0.01642-0.01654	0.02422-0.02438	0.01250-0.01263	0.02414-0.02430	0.01108-0.01118	0.01178-0.01188	0.01541-0.01553
SYR				0.00972-0.00984	0.00822-0.00831	0.00336-0.00343	0.00344-0.00352	0.00762-0.00774	0.01474-0.01488	0.01142-0.01154	0.01916-0.01932	0.00938-0.00951	0.01871-0.01886	0.00728-0.00737	0.00843-0.00854	0.01080-0.01091
ASH					0.00924-0.00934	0.00611-0.00619	0.00534-0.00542	0.00951-0.00963	0.01615-0.01629	0.01087-0.01097	0.01756-0.01772	0.01255-0.01269	0.01740-0.01754	0.01194-0.01204	0.01285-0.01296	0.01752-0.01764
ITJ						0.00508-0.00515	0.00514-0.00521	0.00877-0.00888	0.01455-0.01467	0.01175-0.01186	0.01818-0.01832	0.01287-0.01301	0.01877-0.01891	0.01054-0.01063	0.01099-0.01109	0.01483-0.01494
GRK							0.00048-0.00053	0.00417-0.00427	0.01049-0.01059	0.00730-0.00738	0.01390-0.01402	0.00853-0.00865	0.01444-0.01456	0.00591-0.00598	0.00697-0.00704	0.01052-0.01061
TUR								0.00424-0.00434	0.01054-0.01066	0.00700-0.00710	0.01323-0.01335	0.00835-0.00847	0.01436-0.01448	0.00547-0.00555	0.00612-0.00620	0.00948-0.00957
N_Italian									0.00748-0.00760	0.00190-0.00199	0.00799-0.00811	0.00702-0.00715	0.00826-0.00839	0.01112-0.01124	0.01046-0.01058	0.01744-0.01759
Sardinian										0.00971-0.00981	0.01397-0.01409	0.02039-0.02056	0.02036-0.02052	0.01759-0.01772	0.01779-0.01792	0.02331-0.02346
French											0.00726-0.00737	0.01014-0.01027	0.00554-0.00564	0.01459-0.01469	0.01424-0.01436	0.02100-0.02114
Basque												0.01810-0.01827	0.01426-0.01439	0.02050-0.02064	0.02096-0.02111	0.02733-0.02750
Adygei													0.01334-0.01349	0.01222-0.01235	0.01200-0.01213	0.01769-0.01783
Russian														0.02125-0.02139	0.02187-0.02202	0.02808-0.02824
Palestinian															0.00807-0.00815	0.00825-0.00833
Druze																0.01217-0.01227
Bedouin																

Supplement Table 3. Pairwise *Fst* between major Jewish groups (Middle-East Jews and European/Syrian Jews) and the populations in HGDP panel.

	Middle-East Jews	**European/Syrian Jews**
European/Syrian Jews	**0.00677** (0.00675-0.00680)	
N_Italian	**0.01124** (0.01119-0.01130)	**0.00462** (0.00458-0.00465)
Sardinian	**0.01995** (0.01988-0.02001)	**0.01085** (0.01081-0.01090)
French	**0.01505** (0.01500-0.01510)	**0.00738** (0.00734-0.00741)
Basque	**0.02310** (0.02302-0.02317)	**0.01395** (0.01390-0.01400)
Adygei	**0.01071** (0.01065-0.01077)	**0.00822** (0.00817-0.00827)
Russian	**0.02212** (0.02204-0.02219)	**0.01446** (0.01441-0.01451)
Palestinian	**0.00971** (0.00968-0.00975)	**0.00635** (0.00632-0.00638)
Druze	**0.01021** (0.01017-0.01025)	**0.00701** (0.00699-0.00704)
Bedouin	**0.01457** (0.01452-0.01463)	**0.01101** (0.01097-0.01105)

Supplement table 4. ANOVA of population PCAs average.

PCA1	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR	N_Italian	Sardinian	French	Basque	Adygei	Russian	Palestinian	Druze	Bedouin
IRN		0.1903	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0005	<.0001	<.0001
IRQ			<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0067	<.0001
SYR				<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
ASH					<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.0349	<.0001	<.0001	<.0001	<.0001
ITJ						0.5137	0.6861	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
GRK							0.2918	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
TUR								<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
N_Italian									0.2844	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Sardinian										<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
French											<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Basque												<.0001	<.0001	<.0001	<.0001	<.0001
Adygei													0.3992	<.0001	<.0001	<.0001
Russian														<.0001	<.0001	<.0001
Palestinian															<.0001	<.0001
Druze																<.0001

PCA2	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR	N_Italian	Sardinian	French	Basque	Adygei	Russian	Palestinian	Druze	Bedouin
IRN		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
IRQ			<.0001	<.0001	<.0001	0.003266	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
SYR				0.780547	0.073169	<.0001	0.011957	<.0001	<.0001	<.0001	<.0001	0.002773	<.0001	<.0001	0.757846	<.0001
ASH					0.023892	0.000415	0.002467	<.0001	<.0001	<.0001	<.0001	0.00349	<.0001	<.0001	0.977098	<.0001
ITJ						0.211799	0.400707	0.0006841	<.0001	<.0001	<.0001	<.0001	0.002627	<.0001	0.020277	<.0001
GRK							0.722138	0.015067	<.0001	0.000217	<.0001	<.0001	0.047883	<.0001	0.0003	<.0001
TUR								0.00843921	<.0001	0.000117	<.0001	<.0001	0.027345	<.0001	0.001947	<.0001
N_Italian									<.0001	0.373432	<.0001	<.0001	0.610297	0.1739838	<.0001	<.0001
Sardinian										0.000427	0.317682	<.0001	<.0001	0.00078806	<.0001	<.0001
French											<.0001	<.0001	0.138662	0.6532716	<.0001	<.0001
Basque												<.0001	<.0001	<.0001	<.0001	<.0001
Adygei													<.0001	0.04315468	0.003443	<.0001
Russian														<.0001	<.0001	<.0001
Palestinian															<.0001	<.0001
Druze																<.0001

PCA3	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR	N_Italian	Sardinian	French	Basque	Adygei	Russian	Palestinian	Druze	Bedouin
IRN		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.002932	<.0001	<.0001	<.0001	<.0001
IRQ			<.0001	<.0001	<.0001	<.0001	<.0001	0.58983262	<.0001	<.0001	0.111731		<.0001	<.0001	<.0001	<.0001
SYR				0.074604	<.0001	0.062977	0.787726	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.01334626	0.024763	<.0001
ASH					<.0001	<.0001	0.02597	0.0007077	<.0001	<.0001	<.0001	<.0001	<.0001	0.48075355	0.629081	<.0001
ITJ						<.0001	<.0001	<.0001	0.00508519	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
GRK							0.08374	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
TUR								<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	0.00271076	0.006211	<.0001
N_Italian									<.0001	<.0001	0.059245	<.0001	<.0001	0.00410764	0.002624	<.0001
Sardinian										<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
French											<.0001	<.0001	<.0001	<.0001	<.0001	0.000368
Basque												<.0001	<.0001	<.0001	<.0001	<.0001
Adygei													<.0001	0.01904813	<.0001	<.0001
Russian														<.0001	<.0001	<.0001
Palestinian															0.828882	<.0001
Druze																<.0001

Supplement Table 5. Permutation tests for between-group identity by state (IBS) differences

Populations	*N*	IRN	IRQ	SYR	ASH	ITJ	GRK	TUR
IRN	28		$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$
IRQ	37			$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$
SYR	25				$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$
ASH	34					$1.0X10^{-5}$	$1.0X10^{-5}$	$1.0X10^{-5}$
ITJ	37						$1.0X10^{-5}$	$1.0X10^{-5}$
GRK	42							$1.0X10^{-5}$
TUR	34							

Supplemental table 6 – Gene content for regions of high IBD sharing

Chr	Start	End	Genes
2	193,679,587	196,034,311	--
2	72,175,503	74,001,136	EXOC6B, EMX1, ALMS1, STAMBP, SPR, RAB11FIP5, FBXO41, EGR4, NAT8, DUSP11, ACTG2
8	50,193,145	53,996,125	SNTG1, RB1CC1
9	124,440,201	126,229,508	LHX2, STRBP, CRB2, NEK6, PDCL, GPR21, PSMB7
10	116,184,200	117,850,842	TRUB1
10	2,728,100	3,108,632	--
12	108,046,843	112,554,949	CDV1, TPCN1, ACACB, RAD9B, MYL2, SH2B3, ATXN2, PTPN11, RPH3A, FOXN4, MMAB, MVK, TRPV4, ATP2A2, TECT1, PPP1CC, ALDH2, MAPKAPK5, C12orf8, RPL6, OAS1, OAS3, OAS2, DTX1, RASAL1, DDX54, SLC24A6, LHX5
14	63,153,718	65,330,133	MTHFD1, FUT8, SYNE2, SPTB, MAX, ESR2, ZBTB25, HSPA2, GPX2, FNTB
15	37,023,662	38,192,742	THBS1, GPR176, SRP14, BMF
17	50,619,996	52,424,032	HLF, PCTP, NOG, DGKE, TRIM25, COIL
20	20,304,906	22,377,988	XRN2, NKX2-4, NKX2-2, PAX1