Supporting Information

Bray et al. 10.1073/pnas.1004381107

SI Materials and Methods

Ashkenazi Jewish Ascertainment and Genotyping. Ashkenazi Jewish (AJ) parent-child trios were recruited in the Baltimore, Maryland area by A.E.P., based on a schizophrenia phenotype in the children. All recruitment methods and protocols for collection of clinical data and blood samples were approved by the Johns Hopkins institutional review board, and informed consent was obtained from all individuals. Only the 547 (471 passed quality control) unrelated normal parents were analyzed in this study. For inclusion, children were required to have all four grandparents of Ashkenazi ancestry, thus the parents used in this study had at least two Ashkenazi parents. DNA were genotyped on the Affymetrix 6.0 genomewide SNP array. Genotype calls were made using the birdseed algorithm (v2.0), performed simultaneously with genotyping of HapMap phase 3 CEL files (available from Hapmap.org and Affymetrix). The AJ genotype data have been deposited in NCBI's Gene Expression Omnibus (1) and are accessible through GEO Series accession number GSE23636 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23636).

Genotype Data from European and Worldwide Populations. The number of individuals and genotyped SNPs in each population before and after quality control filters are listed in Dataset S1. The number of individuals listed in this section represents the original raw data before quality-control filtering.

Continental European genotype data were acquired from three sources. All three of these sources used the Affymetrix 500k SNP chip for genotyping. First, we obtained genotyped data from 10 European populations presented by Lao et al. (2). Together, these populations include genotype data for 716 individuals. Second, the genotype data for 500 Germans collected in Kiel, Germany were obtained from the PopGen (3) project at Christian Albrechts University. This is the same Kiel German cohort analyzed by Lao et al. (2) but had to be obtained independently because of user restrictions. Third, we received access to the POPRES (4) population reference sample dataset through the database of Genotypes and Phenotypes (dbGaP), accession number phs000145.v2.p2 (http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v2.p2). Samples were provided to dbGaP by Matthew Nelson, GlaxoSmithKline, and all POPRES genotyping was funded by GlaxoSmithKline (4). Genotypes for 941 individuals from five European populations were extracted based on country of origin for the individual and their parents. Following quality control filters, 1,705 Continental Europeans were used for analysis with 242k SNPs shared with the AJ cohort.

We acquired genotyped data for 1,442 European Americans (EA) that were collected as controls in the Genetic Association Information Network (GAIN) Schizophrenia genomewide association study (5). Samples and associated phenotype data for the GAIN Project Dataset were provided by Pablo Gejman, ENH Research Institute, (5, 6) and the Molecular Genetics of Schizophrenia Collaboration, which received funding from the National Institute of Mental Health. These individuals were genotyped on the Affymetrix 6.0 SNP chip, and genotype data were accessed through the database of genotypes and phenotypes, accession number phs000021.v1.p1 (dbGaP http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000021.v1.p1). Following quality control filters 1,251 EA were used for analysis with 732k SNPs shared with the AJ cohort.

Human Genome Diversity Protject (HGDP) genotype data reported by Li et al. (7) was downloaded at http://hagsc.org/hgdp/files.html. The HGDP data contains genotype data from 938 individuals from 51 population groups worldwide, genotyped on the Illumina 650Y array. Following quality control filters, 883 HGDP individuals were used for analysis with 168k SNPs in common with the AJ cohort.

CEL files from the Affymetrix 6.0 array for the Yoruba from Ibadan, Nigeria (YRI), Utah residents with ancestry from northern and western Europe (CEU), and Toscans in Italy (TSI) were available from the phase III dataset of the HapMap project www.hapmap.org. We downloaded raw CEL files and performed genotype calling simultaneously with the AJ CEL files using Birdseed (v2.0). Following quality control filters, the following individuals were used for analysis: 110 CEU individuals with 732k shared SNPs; 86 TSI with 732k shared SNPs, and 115 YRI with 719k shared SNPs.

Genotype Quality Control Filters. The same quality-control measures were performed for all genotype data from all populations used in the study. Quality-control filters were applied to each population separately unless noted otherwise. First, we removed SNPs that failed the Hardy-Weinberg exact test $P < 0.000001$ and had a call rate <95% across all samples per population. Additionally, individuals with <98% call rate for the passing SNPs were excluded.

Unknown relatedness and population outliers were assessed using PLINK's (v 1.06) (8) pairwise identity-by-state (IBS)/Identity-by-descent (IBD) estimations, IBS_DST and PI_Hat, respectively. If pairs were determined to be too highly related, the individual with the lower call rate was removed from the dataset. Each population was individually analyzed using the genome function. To identify pairs of individuals with evidence of relatedness within their population, we calculated an upper outlier cutoff using five times the interquartile range plus the third quartile (5IQR+Q3). For populations whose upper threshold was above 0.375, we used 0.375 as the upper threshold to exclude first-degree relatedness in the analysis (this primarily affected small HGDP populations from Oceania and Americas). If the median PI_Hat or upper threshold in a population were 0, we set the cutoff to 0.125, to exclude closer than third-degree relations. IBS_DST and PI_Hat were also averaged for each individual for all pairwise comparisons within the population, and an upper threshold was similarly calculated (5IQR+Q3) to exclude individuals too highly related to everyone. We also set a lower cutoff (Q1-5IQR) for individuals not related enough to the population.

The HGDP dataset contains many very small populations, several with 5 to 10 individuals, making it unreasonable to prune for Hardy-Weinberg equilibrium or per SNP call rate within each population. Instead, we grouped populations into regional groups based on Li et al. (7) and used these regions for SNP and individual call-rate and relatedness tests. When performing PLINK's IBS/IBD analysis, each region was evaluated only for relatedness within each individual population using the rel-check parameter with popid substituted for famid.

As a final quality control to exclude outlier individuals from populations, we performed principle component analysis (PCA) using EIGENSOFT (v 3.0) (9, 10) for each population alone and removed outliers across the the first 10 eigenvectors. This process removed individuals greater than six SDs from the mean for five iterations (default). For the HGDP and European cohorts, PCA was performed on each subpopulation independently.

The continental European cohort was treated as a single population to apply SNP quality control filters, but treated as separate subpopulations for IBD and PCA exclusions.

After individual population analysis, we merged the AJ population with the EA cohort and performed PCA to further exclude any Jewish Europeans within the European sample and vice versa. The median value along PC1 was calculated for each population, and individuals outside the median plus or minus three SDs were excluded from further analysis in this study. This process excluded 5 AJ individuals and 32 EA individuals.

Ancestral Clustering. The *frappe* algorithm (v 1.0) (11) was used to determine the ancestral population clustering for the merged AJ and HGDP genotype data. Genotype files were recoded in PLINK (8) and run using the *frappe*_MacOSX version 1.0. Ten-thousand iterations were specified for each run. The AJ population of 471 individuals was divided into three subgroups of 157 individuals and run separately to have an approximately equal sample size relative to the Middle Eastern and European populations in the HGDP cohort. The number of theoretical ancestral populations, K, was defined as K = 7, 3, or 2. Results were plotted using the *Distruct* software (12).

Principle Component Analysis. Principle component analysis was performed using smartpca in the EIGENSOFT (v 3.0) software package (9, 10). To avoid high linkage disequilibrium (LD) between SNPs, we used the kill r2 parameter, with r2thresh set to 0.5. To verify the European ancestry of the European American cohort, we first calculated eigenvectors using only the continental European population with the poplistname parameter and then projected EA, CEU, and TSI populations onto those eigenvectors. We allowed one iteration of outlier removal to exclude approximately eight Netherland individuals that were skewing the results. We restricted the population sizes in the PCA of continental Europeans alone and with the AJ population using the popsizelimit function set to 100. This process helps avoid biases in PCA because of population sizes. When we performed PCA with the AJ and EA populations, we limited the EA population to match the AJ, 471 individuals. We also allowed one iteration of outlier removal, to exclude one Tuscan individual who was an extreme outlier. The continental European populations show that individuals from Finland are very distinct from other Europeans. Because of this finding, we excluded the Finns from the continental European population for the other analysis.

F_{ST} Calculations and Phylogenetic Tree Building. An F_{ST} matrix was calculated using the smartpca function simultaneously to the PCA analysis, by including phylipoutname in the parameters. The phylogenetic tree built using the F_{ST} matrix was created using the FITCH program in the PHYLIP package (v 3.69) (13).

Locus-Specific Admixture Calculations. The LAMPANC algorithm was used from LAMP (v2.3) to calculate the locus-specific admixture given two ancestral populations. Middle East ancestral allele frequencies were taken from Palestinian and Druze populations and European allele frequencies were from either HGDP Europeans (Russians, French, Basque, and Orcadians), Germans from the continental European cohort, or EA. The number of generations since admixture was adjusted to 20, 40, or 80 generations. Recombination rate was set at a fixed rate of 1e-8 per bp per generation and initial alphas were set to 0.7(ME) and 0.3 (Euro). The ancestry for each locus was averaged across all individuals in the population and then the mean was taken across all SNPs.

Genetic Diversity. Heterozygosity (HET) and inbreeding coefficient (F) were calculated based on expected and observed heterozygous calls, with F = (HETexp − HETobs)/(HETexp). Expected HET was calculated from population allele frequencies. The mean HET per SNP and the mean F per individual were calculated within each population. SNPs in high LD, $r^2 >$ 0.5, were pruned out of the dataset using the indep-pairwise 50 5 0.5 setting in PLINK. LD was pruned in each population separately and the remaining shared SNPs between populations were used to calculate the HET and F. For the pairwise IBS test, the AJ individuals were coded as cases and European samples were coded as controls and the ibs-test was run in PLINK with 10,000 permutations. Empirical P values were reported for whether case/case-pairs were less similar to each other compared with control/control-pairs. The AJ population was compared with both the continental European population and the European American cohort. Sample sizes of European groups were also adjusted to be identical to the AJ sample size (471 individuals) and no difference in result was seen.

Linkage Disequilibrium. The r^2 and D' were calculated for SNP pairs within 500 kb of each other in Haploview (v 4.1) (14). European population sizes were set equal to the AJ population (471) to avoid any sample size differences. For the sliding-window analysis, we calculated the average r^2 within a 1.7-Mb window, sliding in 100-kb increments along the genome. For LD decay, the average r^2 was calculated for SNP-pairs based on the distance separating them. Bins were made in increments of 5kb (i.e., 0–5 kb, 5–10 kb, . . ., 495–500 kb). We also directly compared the r^2 for each SNP-pair by tallying the number of pairs that had higher r^2 in one population or the other. The same set of analysis was performed with D' and similar results were observed.

Interpopulation allele-frequency differences, δ, were calculated for 153k shared SNPs between the Middle Eastern (ME, combined Druze and Palestinian) and EA populations. The frequencies were relative to the AJ major allele to ensure that frequencies matched the same allele in both populations. For each pair of the 153k SNPs within 500 kb of each other, the average r^2 in the AJ population was plotted relative to the product of the SNPs' allele frequency differences, $\delta_1\delta_2$. As a control, allele frequency differences between the YRI compared with ME and EA populations were also calculated.

Haplotype Phasing and Frequency Modeling. Haplotypes were phased using the BEAGLE software package (v 3.04) (15). To increase the accuracy of phasing, we used available trios as a reference sample to be phased simultaneously with the target population. Forty-seven trios from the HapMap phase III CEU population were used as the reference for phasing the continental European and EA populations. One hundred fifty-six AJ trios (see AJ ascertainment section, above) were used as a reference to phase the AJ population. Finally, 53 trios from the Hapmap phase III YRI population were used as the reference for phasing the Yoruban population. All Mendelian errors in the trio genotypes were set to missing in PLINK before their use in BEAGLE. BEAGLE parameters were set to –niterations = 20, nsamples = 25. To avoid any sample size differences, only 471 individuals from the continental Europe and European American cohorts were selected to be phased and analyzed in subsequent tests to match the 471 AJ samples.

To model the haplotype structure, the phased haplotypes were used to build a graphical model of haplotype frequency (16). This process was done by invoking the association test without specifying a trait to test. The total number of nodes and edges were summed across all chromosomes as well as calculating the mean nodes and edges per SNP.

Identity-by-Descent. The GERMLINE algorithm (v 1.4.0) (17) was used to estimate segments of IBD between all pairs of individuals from each population. Phased haplotypes from BEAGLE were converted to ped/map format for input in Germline. Minimum IBD segment size was set to 5 Mb, with a 150 SNP seed (-bits 150)

and all other default parameters. Mean segments per pair was calculated based on the total segments divided by the total possible pairwise comparisons. Similarly, the mean shared genome per pair was calculated based on the sum of the segment lengths divided by the total possible pairwise comparisons. The percentage of pairs within each population that shared each SNP across the genome was calculated as the number of times it was within a shared segment divided by the total possible pairwise comparisons. Segments of IBD were additionally filtered in the AJ and EA populations by using a 1-Mb nonoverlapping sliding window across the genome to identify low-density SNP regions (<100 SNPs/Mb). These regions were then excised from any overlapping IBD segments and only segments that had at least 3 Mb of high density SNP coverage were kept. Using the filtered data we also analyzed the IBD length decay and maximum-length distribution in the AJ population. For the IBD decay we calculated the total number of segments in each 1-Mb interval and plotted the number of segments versus the length. For the max length distribution, the largest segment of IBD between each pair of individuals was identified and the total number within each 1-Mb interval was plotted versus the length.

Positive Selection. The integrated haplotype score (iHS) and cross population extended haplotype homozygosity (XP-EHH) were implemented according to the methods previously found to give the greatest power to detect selected regions (18–20). Scripts to calculate iHS and XP-EHH scores were available from the HGDP selection browser sponsored by the Pritchard laboratory (http://hgdp.uchicago.edu/). Haplotype data, only using SNPS with MAF > 0.05, were used to calculate unstandardized iHS scores for each population. Chimpanzee ancestral alleles for SNPs on the Affymetrix 6.0 array were taken from SNP annotations compiled by the P. Sullivan laboratory (University of North Carolina, Chapel Hill) and are available via PLINK. SNP annotations were created using the TAMAL (21) database based chiefly on the University of California Santa Cruz genome browser files, HapMap, and dbSNP. To normalize the iHS score with SNPs of similar allele frequency, we first calculated the mean iHS and SD for SNPs in 20 equally sized allele-frequency bins. We then normalized the iHS scores by subtracting the mean and dividing by the SD, giving a mean of 0 and SD of 1. To identify regions that had the highest proportion of SNPs with extreme iHS scores, we used nonoverlapping 40-SNP windows across the genome and calculated the fraction of SNPs within each window that had an iHS > 2. Idiograms of iHS windows were created with Idiographica (22).

The XP-EHH was calculated by comparing the AJ and EA populations directly. XP-EHH scores were normalized by subtracting the mean and dividing by the SD of all scores. We again used nonoverlapping 40-SNP windows across the genome and ranked the windows based on the maximum |XP-EHH|. The direction of selection was determined by the sign of the XP-EHH, in this case negative, indicating selection in the European population, and positive, indicating selection in the AJ population.

Statistical Analysis. When comparing the means between populations (i.e., heterozygosity) we used a standard two-tailed *t* test to calculate the significance. In our comparison of LD, we compared the r^2 of each SNP pair between populations and tallied the number of pairs that were higher or lower. The significance of this binomial test was calculated using the number of successes (higher) and failures (lower), with a probability of either 0.5. The empirical *P* value of the IBS test was determined in PLINK by permuting case(AJ)/control(European) labels 10,000 times and calculating IBS metrics within and between groups at each permutation.

1. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30:207–210.
2. Lao O, et al. (2008) Correlation between genetic and geographic structure in Europe. *Curr Biol* 18:1241–1248.
3. Krawczak M, et al. (2006) PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet* 9:55–61.
4. Nelson MR, et al. (2008) The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 83:347–358.
5. Sanders AR, et al. (2008) No significant association of 14 candidate genes with schizophrenia in a large European ancestry sample: implications for psychiatric genetics. *Am J Psychiatry* 165:497–506.
6. Manolio TA, et al.; GAIN Collaborative Research Group Collaborative Association Study of Psoriasis International Multi-Center ADHD Genetics Project Molecular Genetics of Schizophrenia Collaboration Bipolar Genome Study Major Depression Stage 1 Genome-wide Association in Population-Based Samples Study Genetics of Kidneys in Diabetes (GoKinD) Study (2007) New models of collaboration in genome-wide association studies: The Genetic Association Information Network. *Nat Genet* 39:1045–1051.
7. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
8. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
9. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.
10. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909.
11. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical and study design considerations. *Genet Epidemiol* 28:289–301.
12. Rosenberg NA (2004) *Distruct*: A program for the graphical display of population structure. *Mol Ecol Notes* 4:137–138.
13. Felsenstein J (2009) PHYLIP (Phylogeny Inference Package) version 3.69 Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, WA.
14. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* 21:263–265.
15. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–223.
16. Browning SR (2006) Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 78:903–913.
17. Gusev A, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19:318–326.
18. Pickrell JK, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826–837.
19. Sabeti PC, et al.; (2007) International HapMap Consortium (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918.
20. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.
21. Hemminger BM, Saelim B, Sullivan PF (2006) TAMAL: An integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics* 22:626–627.
22. Kin T, Ono Y (2007) Idiographica: A general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics* 23:2945–2946.

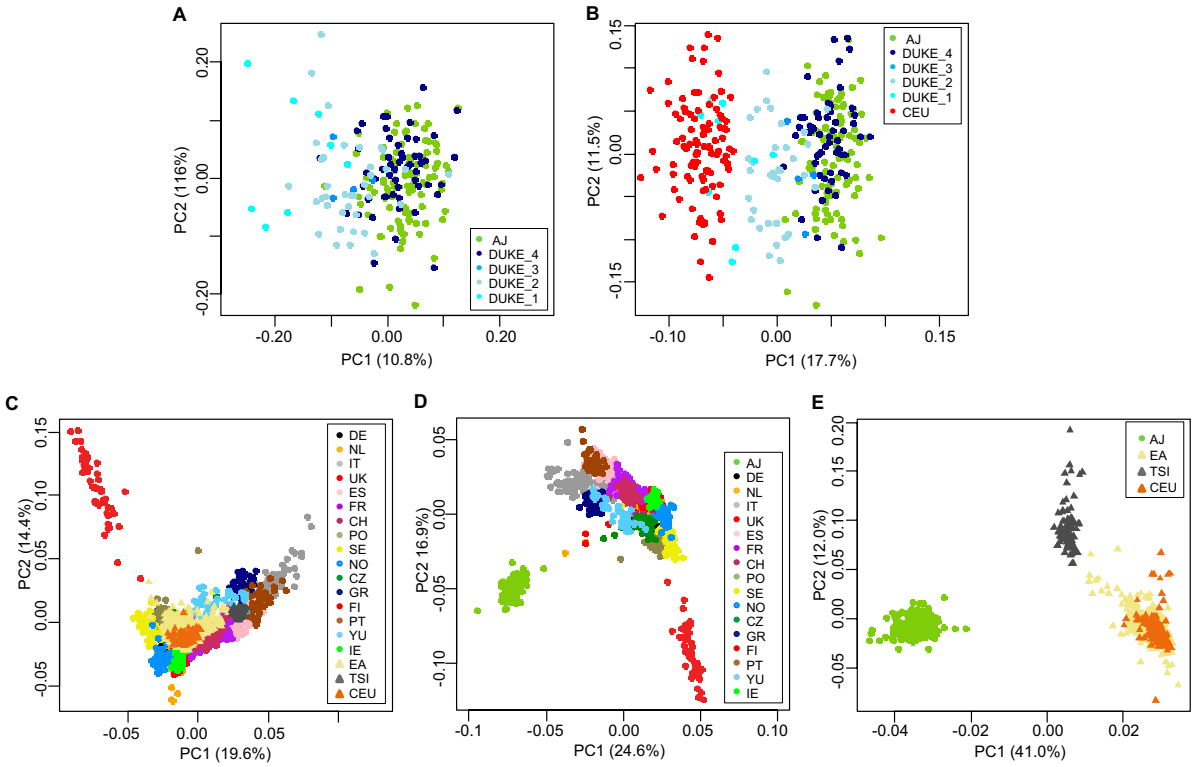
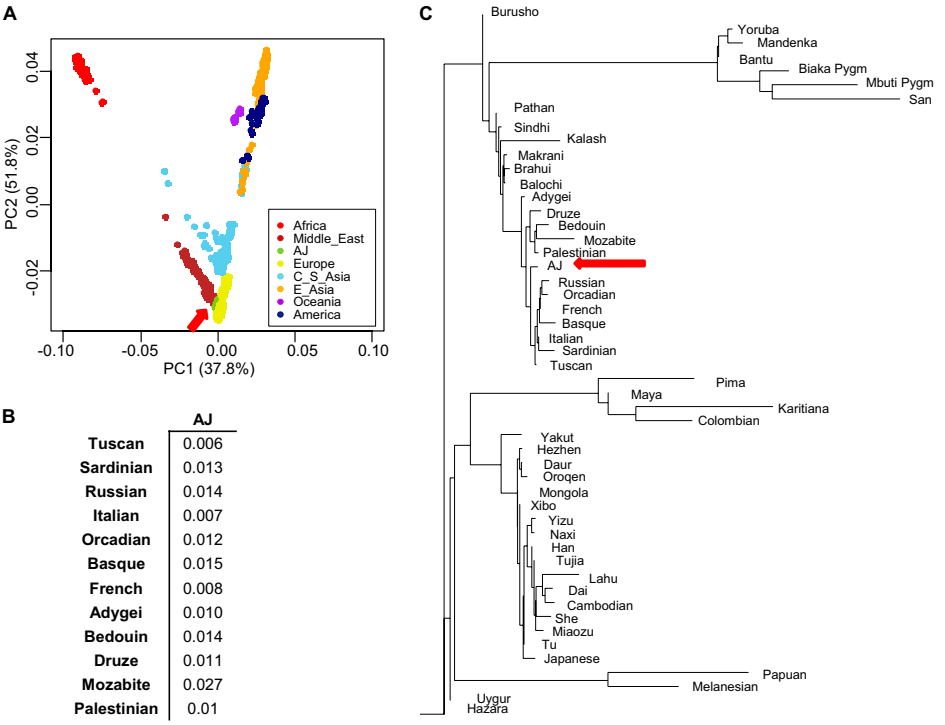


Fig. S1. Principal component analysis of AJ and European populations. (*A*) PCA of 100 AJ individuals from our cohort compared with the Need et al. (1) Jewish cohort from Duke, using 3,223 overlapping SNPs. DUE_4 refers to four reported Jewish grandparents, DUE_3 refers to three reported Jewish grandparents, and so forth. (*B*) Same as *A*, with the addition of the CEPH; Utah residents with ancestry from northern and western Europe (CEU). (*C*) PCA of continental European populations with no Ashkenazi Jews included. PC1 and PC2 roughly separate continental European subpopulations by latitude and longitude, respectively. Individuals from the EA, CEU, and Toscans in Italy (TSI) were projected onto the principal components of the continental Europeans to confirm the accuracy and relative location of their origin. Abbreviations for European countries are according to standard nomenclature and can be found in Dataset S1. (*D*) PCA with the AJ population combined with the continental European populations. (*E*) PCA with the AJ population combined with the EA, CEU, and TSI populations. The percentage of variance explained by each principle component is shown along the axis in parenthesis.

1. 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:R7.



B

	AJ
Tuscan	0.006
Sardinian	0.013
Russian	0.014
Italian	0.007
Orcadian	0.012
Basque	0.015
French	0.008
Adygei	0.010
Bedouin	0.014
Druze	0.011
Mozabite	0.027
Palestinian	0.01

Fig. S2. PCA, F_{ST} and phylogenetic tree of AJ and HGDP populations. PCA and calculation of the fixation index, F_{ST}, were performed on the merged genotype data from the AJ and HGDP populations (168k overlapping SNPs). The AJ population was divided into three random subgroups containing 157 individuals to better match the population size of the ME and European populations in the HGDP dataset. The data shown here represent one subgroup, and all three had similar results. (*A*) PCA of the AJ population combined with all populations of the HGDP dataset, colored by regional origin shown in legend. The red arrow highlights the AJ population. (*B*) The F_{ST} matrix was calculated concurrently with the PCA using the AJ and all 50 HGDP populations. Only the European and ME populations compared with the AJ population are shown here. (*C*) An unrooted phylogenetic tree built using the F_{ST} calculated between the AJ and all HGDP populations. The red arrow again highlights the AJ position.

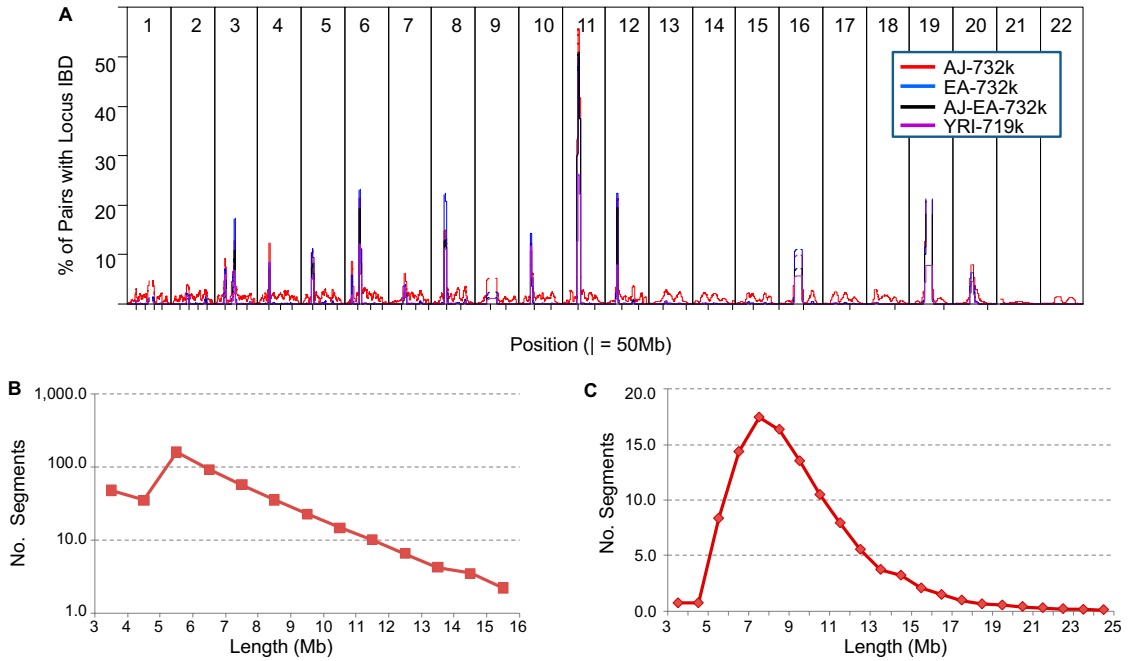


Fig. S3. Shared segments of IBD across the genome. All pairs of individuals within each population were analyzed for genomic segments of IBD ≥ 5 Mb. (*A*) The percentage of pairs in each population that share a given locus of IBD is plotted against the physical position across the autosomes. In addition to the AJ and EA populations, the outgroup population, the Yoruba in Ibadan, Nigeria (YRI) was analyzed, as were pairs consisting of one Ashkenazi Jew with one European American, AJ-EA. (*B*) IBD length decay in the AJ population was plotted for IBD segments after filtering for low-density SNP regions. The number of segments (in thousands) in each 1-Mb interval is plotted in log scale. (*C*) The maximum IBD length distribution for the AJ population after filtering for low-density SNP regions. The longest IBD segment for each pair of individuals was identified and the number (in thousands) in each 1-Mb interval is plotted.

Table S1. Direct paired comparison of linkage disequilibrium in Ashkenazi and European populations

Distance between SNPs	AJ > EA	EA > AJ	Ratio AJ:EA	AJ > Euro	Euro > AJ	Ratio AJ:Euro	Sign test *P* value <
< 200kb	19,110,423	16,011,531	1.19	2,177,701	1,744,618	1.25	2.20E-16
≥ 200kb < 500kb	28,883,192	18,000,378	1.60	3,176,496	1,990,443	1.60	2.20E-16
< 500kb	47,993,615	34,011,909	1.41	5,354,197	3,735,061	1.43	2.20E-16
Nearest neighbor	183,053	178,393	1.03	61,164	55,908	1.09	9.22E-15

The r^2 value for each pair of SNPs across the genome was directly compared between populations. The number of SNP pairs with r^2 greater in AJ, Euro, or EA is reported for different distances between SNPs. The European population groups were limited to 471 individuals to ensure no sample size effects.

Table S2. Haplotype diversity among AJ and Europeans

Pop_SNPs	Nodes	Edges	Edges/node	Nodes/level	Edges/level
AJ_242K	572,764	864,772	1.51 ± 0.50*	2.36 ± 1.08*	3.56 ± 1.47*
Euro_242K	567,779	893,412	1.57 ± 0.49	2.34 ± 1.06	3.68 ± 1.54
AJ_732K	2,536,896	3,546,844	1.40 ± 0.49*	3.46 ± 1.28*	4.84 ± 1.69*
EA_732K	2,509,126	3,685,217	1.47 ± 0.50	3.43 ± 1.28	5.03 ± 1.78

The BEAGLE[27] haplotype modeling algorithm was used to represent haplotype diversity in terms of nodes and edges at each marker level across the genome (±SD).
**t* test P < 1e-50.

Table S3. Segments of IBD

Pop_SNPs	Ind	Pairwise comparisons	Segments/pair	Size (Mb)/segment	Total (Mb)/pair
AJ_242k	471	110,685	5.41 ± 2.32*	8.40 ± 4.23*	45.45 ± 21.70*
Euro_242k	471	110,685	0.77 ± 0.86	8.50 ± 2.98	6.53 ± 10.04
AJ_732k	471	110,685	5.94 ± 2.33*	8.30 ± 4.33*	49.29 ± 21.73*
EA_732k	471	110,685	2.53 ± 1.43	8.37 ± 3.73	21.19 ± 13.24
YRI_719k	115	6,555	1.36 ± 1.10	8.23 ± 3.84	11.16 ± 10.23
AJ-EA_732k	942	221,841	1.93 ± 1.23	8.21 ± 3.34	15.81 ± 11.01
AJ_732k_f	471	110,685	4.51 ± 2.10*	6.69 ± 2.62*	30.17 ± 15.20*
EA_732k_f	471	110,685	0.95 ± 0.93	4.68 ± 1.76	4.47 ± 4.75

All pairs of individuals within each population were analyzed for genomic segments of IBD ≥ 5 Mb. The number of shared segments, the size per segment, and total genome shared per pair of individuals is reported for each population, ± SD. An out-group population, the YRI were analyzed, as were pairs that consisted of one Ashkenazi Jew with one European American, AJ-EA. Additionally, the IBD segments in the AJ and EA populations were also filtered (_f) to remove regions with low-density SNP coverage (i.e., at centromeres).
*t test, $P < 1e\text{-}10$.

Table S4. Top iHS regions of selection for AJ and EA populations

Chr: window (Mb hg18)	Pop	Fraction of SNPs w/|iHS| > 2	Max |iHS|	Genes (number)	Previously implicated (within top reported regions)
1: 31.55–31.76	AJ	0.60	3.84	*FABP3, SERINC2* (3)	
2: 74.33–74.83	AJ/EA	0.65/0.65	3.36/3.36	*SLC4A5, DCTN1, RTKN, HTRA2,DOK1, SEMA4F* (24)	(1)
2: 134.94–137	EA	0.63	5.02	*CCNT2, RAB3GAP1, LCT, CXCR4* (12)	(1, 2)
2: 197.13–197.77	EA	0.58	4.10	*PGAP1* (6)	(1–3)
4: 33.4–34.68	AJ/EA	0.65/0.65	4.40/4.40	None (0)	(1–3)
4: 43.03–43.37	EA	0.65	3.89	None (0)	
4: 148.23–148.68	EA	0.54	3.48	*EDNRA* (1)	(1–3)
5: 111.25–111.65	AJ	0.68	6.33	*EPB41L4A* (4)	
5: 112.37–112.58	AJ/EA	0.59/0.78	3.90/3.89	*DCP2, MCC* (2)	(2, 3)
5: 142.07–142.49	AJ	0.59	4.38	*ARHGAP26* (1)	(1, 3)
6: 22.51–22.63	AJ/EA	0.78/0.68	4.41/4.59	None (0)	
6: 30.17–30.46	EA	0.61	5.30	*TRIM* family (9)	(1, 3)
6: 30.87–31.62	AJ/EA	0.48/0.5	6.35/8.62	*HLA-C, HLA-B* (25)	
6: 92.83–93.05	AJ	0.65	5.40	None (0)	
7: 133.41–133.68	AJ	0.78	4.11	*LRGU.K., SLC35B4* (2)	
8: 51.93–53.12	EA	0.55	4.39	*PXDNL, PCMTD1* (2)	(1–3)
8: 139.58–139.87	AJ	0.78	4.05	*COL22A1* (1)	(1, 3)
8: 140–140.16	AJ	0.78	3.39	None (0)	(3)
10: 83.82–84.14	AJ	0.50	3.93	*NRG3* (1)	(2, 3)
12: 87.26–88	AJ	0.54	3.60	*KITLG* (1)	(1)
12: 109.87–110.24	EA	0.88	3.40	*CUX2* (1)	(2, 3)
12: 109.77–111.72	AJ	0.73	5.18	*CUX2, ATXN2, ALDH2, MAPKAPK5, TRAFD1, BRAP, PTPN11* (18)	(2, 3)
13: 67.94–68.37	AJ	0.49	5.10	*MIR548H4* (1)	
14: 65.57–66.32	EA	0.53	3.87	*GPHN* (2)	
14: 101.42–101.88	AJ/EA	0.6/0.58	3.52/3.51	*PPP2R5C, DYNC1H1, HSP90AA1, RAGE* (6)	(1)
15: 27.02–27.22	AJ	0.73	4.01	*APBA2* (2)	(3)
15: 82.03–82.26	EA	0.63	3.82	*SH3GL3, ADAMTSL3* (2)	
15: 91.83–91.95	AJ	0.60	3.65	None (0)	(3)
16: 78.35–78.46	EA	0.70	3.87	None (0)	(1–3)
17: 40.91–41.59	AJ/EA	0.55/0.62	5.79/7.69	*CRHR1, MAPT* (13) - known inversion	(1)
18: 7.45–7.66	EA	0.75	4.60	*PTPRM* (1)	(1–3)
20: 33.35–34.29	EA	0.79	4.43	*GDF5, CEP250, SPAG4, NFS1, EPB41L1* (17)	(1, 2)
20: 39.46–39.72	EA	0.58	3.81	*CHD6* (1)	

The top 20 iHS regions in the AJ and EA populations according to the fraction of SNPs with |iHS| >2 in nonoverlapping 40-SNP windows across the genome. Consecutive SNP windows that were in the top 1% were combined to identify a single region. Regions that overlapped in both populations were merged in the table and are denoted AJ/EA in the Pop column. A few genes for each region are listed, with total number in parenthesis. Regions that overlap with previously identified regions of selection in European or Middle Eastern populations are indicated by references.

1. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.
2. Frazer KA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
3. Pickrell JK, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826–837.

Table S5. iHS regions showing differential selective strength

Chr: window (Mb hg18)	Pop	Fraction of SNPs |iHS| > 2	Max |iHS|	Genes (number)
1: 234.76–234.94	EA	0.49	3.27	*LGALS8, HEATR1, ACTN2* (3)
2: 134.94–137.12	EA	0.63	5.02	*CCNT2, RAB3GAP1, LCT, CXCR4* (12)
2: 163.97–164.16	AJ	0.58	4.40	None (0)
2: 197.13–197.77	EA	0.58	4.10	*PGAP1* (6)
4: 41.03–41.23	EA	0.58	3.38	*LIMCH1* (1)
4: 43.03–43.37	EA	0.65	3.89	None (0)
4: 81.54–82.32	EA	0.53	3.30	*BMP3, PRKG2* (3)
5: 142.07–142.28	AJ	0.73	4.38	*ARHGAP26* (1)
5: 145.95–146.11	EA	0.58	4.24	*PPP2R2B* (1)
6: 29.81–29.89	AJ	0.55	4.32	*IFITM4P* (3)
6: 30.87–31.13	AJ	0.54	6.19	*DDR1, DPCR1, MUC21* (6)
6: 105.56–105.85	EA	0.58	2.92	*LIN28B, BVES, POPDC3, PREP* (4)
8: 51.93–52.44	EA	0.55	3.86	*PXDNL1* (1)
9: 3.96–4.11	AJ	0.49	5.97	*GLIS3* (1)
9: 125.08–125.81	EA	0.48	3.23	*CRB2, DNND1A* (3)
10: 68.68–69.12	AJ	0.55	3.25	*CTNNA3* (1)
10: 83.82–84.14	AJ	0.50	3.93	*NRG3* (1)
12: 110.65–111.72	AJ	0.56	3.95	*ALDH2, MAPKAPK5, TRAFD1, PTPN11* (11)
15: 27.02–27.22	AJ	0.73	4.01	*APBA2* (1)
15: 91.83–91.95	AJ	0.60	3.65	*AK094352* (1)

Regions of differential selective strength for the AJ and EA populations are defined as regions in the top1% of iHS hits in one population but not in the top 1% of the other. Pop indicates which population shows stronger selection. The top 10 regions for each population are listed. Consecutive SNP windows that showed differential selection were combined to identify a single larger region. A few genes in the region are listed, with the total number in parentheses.

Other Supporting Information Files

Dataset 1 (XLSX)