High-resolution inference of genetic relationships among Jewish populations

Naama M Kopelman, Lewi Stone, Dena G Hernandez, Dov Gefel, Andrew B Singleton, Evelyne Heyer, Marcus W Feldman, Jossi Hillel, Noah A Rosenberg

SUPPLEMENTARY MATERIALS AND METHODS

Samples. We obtained 535 samples from people representing 32 populations, 29 Jewish and 3 non-Jewish (Karaite, Palestinian, Tajik). Samples were obtained from four sources (**Table S7**). We collected new samples at (1) Barzilai Medical Center in Ashkelon, Israel, at (2) a Karaite community center in Ashkelon, Israel. We then merged them with previously collected samples from the (3) the National Laboratory for the Genetics of Israeli Populations (NLGIP) , and (4) the laboratory of E. Heyer. For samples collected specifically for this study, we excluded subjects who did not have all four grandparents originating from the same Jewish community. The same criteria applied for inclusion in NLGIP. Informed consent was obtained, and ethics approvals were provided by the Barzilai Medical Center and the University of Michigan.

For the 91 samples collected for the current study (third column of **Table S7**), DNA was extracted following the Gentra purification kit protocol. Genotyping for the 535 samples was performed at the National Institute on Aging using the Illumina Human660WQuadv1 BeadChip (Illumina Inc., San Diego, CA). Among 535 samples sent for genotyping, 79 samples were excluded due to low DNA concentration. Thus, the initial sample set after genotyping had 456 individuals.

After genotyping, the data consisted of 456 samples, each with genotypes at 657,366 markers. The markers included 640,663 autosomal loci, 16,509 X-chromosomal loci, 44 Y-chromosomal loci, 135 mitochondrial loci, and 15 loci on the pseudoautosomal part of the X chromosome (XY loci).

Sex Assignments. To verify the reported sex information for the samples collected, we used the X-chromosomal loci. For each individual, we computed the fraction of X-chromosomal loci that were heterozygous as a function of the fraction of data missing on the X chromosome (**Fig. S2**). For this computation, we used the initial 456 samples and the 16,509 X-chromosomal SNPs. For the purpose of sex assignment, missing data rate was calculated as the fraction of SNPs for which the two alleles were missing. X-chromosomal heterozygosity was calculated using sites with non-missing data.

The mean X-chromosomal heterozygosity for samples with a prior male assignment was 0.0052, and it was 0.3090 for samples with a prior female assignment (**Fig. S2**). We identified 7 samples whose X-chromosomal heterozygosities were outliers with respect to other samples with the same sex assignment. These outliers included 3 male samples with mean X-chromosomal heterozygosity of 0.3458 and minimum 0.3197, and 4 female samples with mean X heterozygosity 9.318×10^{-4}, and maximum 0.0017. For 19 samples, no initial sex assignment was provided. After

correcting the 7 outliers and assigning the 19 samples with no initial assignment, the maximal heterozygosity of males was 0.0111, and the minimal heterozygosity of females was 0.2263.

At this point, apparent heterozygotes in males for X and Y-chromosomal loci were recorded as missing data, as were heterozygotes at mitochondrial SNPs in all individuals and non-missing Y-chromosomal genotypes in females.

SNP Quality Control. We next implemented additional procedures to exclude low-quality SNPs.

SNPs with more than two alleles: The 456 samples were scanned to verify that no SNPs had more than two distinct alleles detected in the full set of individuals. No SNPs had this problem.

Monomorphic SNPs: The 456 samples were scanned to identify SNPs monomorphic across the entire set of individuals. We identified 2,600 monomorphic SNPs—2,445 autosomal, 133 X-chromosomal, 3 Y-chromosomal, and 19 mitochondrial—and removed them from the analysis.

SNPs with more than 10% missing data: Next, we scanned the 456 samples to identify SNPs that had >10% missing data. We identified and removed 96,653 such SNPs—93,469 autosomal, 3,141 X-chromosomal, 38 Y-chromosomal, and 5 mitochondrial. For X-chromosomal SNPs, the missing data fraction was calculated as the number of alleles missing, accounting for two alleles in females and one in males. The Y-chromosomal missing data calculation assumed one allele for males only.

Summary: Based on these steps, we removed 99,253 from the 657,366 SNPs in the initial dataset, leaving 558,113 SNPs (544,749 autosomal, 13,235 X-chromosomal, 3 Y-chromosomal, 111 mitochondrial, and 15 pseudoautosomal).

Duplicates and Relatives. Identification of duplicates and pairs of close relatives was performed using identity-in-state (IIS) allele sharing combined with likelihood inference, as Rosenberg (2006). First, for each pair of individuals, we determined the proportion of the autosomal SNPs at which the pair shared 0, 1, and 2 alleles identical by state. Of the 544,749 autosomal SNPs that passed quality control, only SNPs for which neither individual in a pair was missing genotypes were included in the calculation for that pair. With this procedure, we detected 7 duplicate pairs among NLGIP Bulgarian, Ethiopian, Hungarian, and Kurdish Jewish samples (**Table S8**) and 10 pairs of apparent relatives among the Palestinian samples and the Cochin, Libyan, and Polish Jewish samples (**Fig. S3**).

After identification of apparent relative pairs from IIS ratios, the four populations in which these pairs were detected were then screened for close relatives using *RELPAIR* (Boehnke & Cox 1997; Epstein et al. 2000). This analysis searched for pairs with a relationship closer than first cousins. We used an estimated genotyping error rate of 0.001 and a critical value of 100 for the *RELPAIR* likelihood ratio computation. In each population, *RELPAIR* was applied with count estimates of allele frequencies in that population. Because of a limit on the number of SNPs allowed by *RELPAIR*, a subset of the 544,749 autosomal SNPs was used. Separately for each population, we identified polymorphic SNPs and sorted them, first by chromosome, and within chromosomes, by physical location. Genetic map

positions used for arranging the SNPs were determined by interpolation on the Rutgers combined linkage-physical map (Matise et al. 2007). Vector positions were numbered from 0. The maximal number of SNPs allowed by *RELPAIR* (9,999) was chosen, with SNPs evenly spaced in the vector. Denoting $s=\lfloor n_{pop}/9999 \rfloor$, where n_{pop} is the number of polymorphic SNPs for population *pop*, the final SNP set for a population included 9,999 SNPs numbered $s(j\text{-}1)$ with j ranging from 1 to 9,999.

One parent-offspring (PO) pair was inferred for the Polish Jewish population, one pair of full siblings (FS) was inferred for the Libyan Jewish population, and one avuncular (AV) pair was inferred for Palestinians. Six FS and one AV pairs were inferred for the Cochin Jewish population. As IIS analysis and *RELPAIR* agreed on the specific pairs, we eliminated samples to produce a set of individuals with no duplicates and no pairs related at a level closer than first cousins. We thus omitted seven duplicates and three samples of unknown origin (**Tables S7** and **S8**), and based on the relatedness analysis, one Palestinian sample, one Polish Jewish sample, one Libyan Jewish sample, and five Cochin Jewish samples (**Table S9**). Note that the number of samples removed due to relatedness is less than the number of relative pairs, as some individuals appeared in more than one relative pair. When an arbitrary decision was required about which individual in a pair to exclude, the individual with more missing data was discarded. **Table S7** reports the numbers of samples omitted due to quality issues, duplication, and relatedness. The total number remaining at this stage was 438.

Hardy-Weinberg Disequilibrium. From the sample of 438 individuals, two population groupings with relatively low levels of population structure were constructed in order to perform tests for Hardy Weinberg equilibrium at each SNP: Ashkenazi (1 Belorussian, 10 Czech, 12 German, 5 Latvian, 16 Lithuanian, 20 Polish, 17 Romanian, 26 Russian, and 18 Ukrainian; 125 total individuals), and Mizrahi (21 Georgian, 20 Iranian, 25 Iraqi, 8 Kurdish; 74 total individuals).

A chi-square test of the null hypothesis of Hardy-Weinberg equilibrium was performed in each of the two population groups, taking into account the Yates continuity correction (Weir 1996). For X-chromosomal SNPs, males were included in the calculation of allele frequencies but not in the test. Only SNPs with at least 4 copies of the minor allele in both groups were considered as candidates for exclusion. Among such SNPs, those SNPs that had either or both of the following properties were identified: (1) the chi-square test statistic exceeded 19.51142 ($p<10^{-5}$, 1 df) in either of the two groups; (2) the test statistic exceeded 6.634897 ($p<10^{-2}$, 1 df) in both groups. Using these criteria, which follow those of Jakobsson et al. (2008), we excluded 341 SNPs (340 autosomal, 1 X- chromosomal).

Summary of Quality Control Steps. **Figure S4** summarizes the preprocessing, indicating numbers of individuals and SNPs omitted from the dataset in each step of the quality control. The final number of individuals included in the analysis was 438, and the final number of SNPs for analysis was

557,772: 544,409 autosomal and 13,363 non-autosomal (13,234 X-chromosomal, 3 Y-chromosomal, 111 mitochondrial, and 15 XY). In the following analyses, only autosomal SNPs were used.

Overview of the Data Merging Process. We combined unphased genotype data from HGDP, HapMap, and Behar et al. (2010) together with the data generated as part of this study.

Unphased Genotypes from HGDP and HapMap. Using procedures of Pemberton et al. (2010, 2012), a merged HGDP and HapMap III unphased dataset was created, containing 938 unrelated individuals from the H952 HGDP subset (Rosenberg 2006) and 1,117 unrelated individuals from the HAP1117 HapMap subset (Pemberton et al. 2010). For HGDP (Li et al. 2008), genotypes for 644,258 autosomal SNPs were used. Quality control was performed as in Pemberton et al. (2010, 2012). After quality control, the final HGDP dataset contained 642,999 SNPs. Monomorphic SNPs (53) and SNPs with >10% missing genotypes (339) in the 938 individuals were removed. A further 88 SNPs with sample size <5 alleles in at least one of the 53 HGDP populations were omitted, as were 779 SNPs with Hardy-Weinberg disequilibrium in at least one of two groups with low levels of populations structure—a Middle Eastern group (Bedouin, Druze, Palestinian; 133 individuals), and a sub-Saharan African group (Bantu from southern Africa, Bantu from Kenya, Mandenka, Yoruba; 62 individuals). The Yates-corrected chi-square test followed the same criteria as above (Weir 1996).

For HapMap (International HapMap 3 Consortium 2010), unphased genotypes were available at 1,423,833 autosomal SNPs. After quality control, the final HapMap dataset contained 1,405,599 SNPs. We removed 424 SNPs monomorphic in the 1,117 individuals. A further 17,810 SNPs were excluded because of Hardy-Weinberg disequilibrium, following the criteria of Pemberton et al. (2010). SNPs were excluded if they had a Yates-corrected chi-square statistic >19.51142 ($p<10^{-5}$, 1 df) in at least one population or a statistic >6.634897 ($p<10^{-2}$, 1 df) in at least two populations (taking into account only populations in which there were at least four copies of the minor allele).

The combined HGDP–HapMap set consisted of 2,055 individuals from 64 populations and 590,461 autosomal SNPs that the two datasets shared in common. Where genotypes had opposite strands, HGDP data was converted to match HapMap data. **Figure S5** summarizes the quality control and merging processes performed on the HGDP and HapMap datasets.

We next assembled a dataset containing the combined HGDP and HapMap data and our new data. This set initially consisted of 2,493 samples at 488,956 autosomal, 4,314 X, and 13 XY SNPs that the two datasets shared in common. The combined set was scanned for duplicates and relative pairs using identity-by-state allele sharing. Seven duplicate and three relative pairs were found among the combined Palestinian sample, each involving one sample from our data and a second sample from HGDP data. Genotypes of duplicate pairs were compared to ensure that there were no data-source-specific biases; none of the SNPs had differing alleles between duplicate samples for more than two duplicated pairs. Following this check, we removed the 7 duplicates and 2 samples with relatives in

the combined set (one sample appeared in two of the relative pairs). In these cases, we removed the sample from our new data and kept the HGDP sample, leaving 429 samples from the new data,. Thus, the final merged set of our data with HGDP and HapMap contained 2,484 samples.

Unphased Genotypes from Behar et al. (2010). Quality control for genotypes from Behar et al. followed a similar procedure to that used for new samples. We obtained 478 samples with genotypes at 544,485 SNPs from http://www.evolutsioon.ut.ee/MAIT/jew_data/ (downloaded September 2011). After discarding three samples excluded by Behar et al., we screened the 475 remaining samples to verify reported sex information, using heterozygosity at the 13,032 available X-chromosomal SNPs. Missing data rate was calculated as the fraction of SNPs for which the two alleles were missing. For one sample with unreported sex, sex was determined from X-chromosomal heterozygosity. Apparent heterozygotes in males for X- and Y-chromosomal loci were then coded as missing, as were mitochondrial heterozygotes in all individuals and non-missing Y-chromosomal genotypes in females.

We excluded 433 monomorphic SNPs: 386 autosomal, 34 X-chromosomal, 1 Y-chromosomal, and 12 mitochondrial. In addition, we excluded 8 autosomal SNPs with >10% missing data. Identification of relative pairs was performed using IIS allele sharing. We identified 3 duplicate pairs and 6 pairs of apparent relatives among the Ethiopian Jews, Hungarians, Iranians, Iraqi Jews, Samaritans, South Indians, and Yemenites, removing one sample from each pair based on missing data rate. We then constructed two population groupings to perform chi-square tests for Hardy-Weinberg equilibrium at each SNP. The two groupings contained two samples from the Middle East and Central and South Asia, one containing 20 Jordanians, 7 Lebanese, 20 Saudi Arabians, and 16 Syrians, for a total of 63 individuals, and the other containing 19 Armenians, 20 Georgians, 20 Iranians, 18 Lezgins, and 19 Turkish, for a total of 96 individuals. A Yates-corrected chi-square test of the null hypothesis of Hardy-Weinberg equilibrium was performed in each group. Only SNPs with at least 4 copies of the minor allele in both population groups were considered as candidates for exclusion. SNPs that had either or both of the following properties were identified: (1) the chi-square statistic exceeded 19.51142 ($p<10^{-5}$, 1 df) in either of the two groups; (2) the chi-square statistic exceeded 6.634897 ($p<10^{-2}$, 1 df) in both groups. Using these criteria, we excluded 1,697 SNPs: 142 autosomal and 1,555 X-chromosomal. **Figure S6** summarizes the preprocessing steps on the raw data of Behar et al., indicating the numbers of individuals and SNPs omitted from the dataset at the various steps. The final number of individuals following the preprocessing was 466, and the final number of SNPs before further merging was 542,347: 530,779 autosomal and 11,568 X-chromosomal.

We scanned a merged dataset containing our data and that of Behar et al. for duplicates and relatives using IIS allele-sharing, using 486,592 autosomal SNPs that all the datasets shared in common. We identified 15 duplicates and 1 relative pair among the combined Bulgarian Jewish, Ethiopian Jewish, Hungarian Jewish, Indian Jewish (Cochin), Iranian Jewish, Iraqi Jewish, and Yemenite Jewish samples. Each duplicate pair involved one sample from our data and a second

sample from Behar et al.; one sample originating from Behar et al. was removed from each pair. Thus, 450 samples at 542,347 remained from the data of Behar et al. **Table S10** lists the samples used from Behar et al. according to population source.

Final Combined Data Set. We assembled a dataset with 2,934 individuals (938 HGDP-CEPH, 1,117 HapMap, 450 Behar et al., 429 new samples) from 122 populations. This population set included 32 newly sampled populations; 53 HDGP populations including one, Palestinians, that overlapped with the new data; 11 HapMap populations; 48 Behar et al. populations including 21 populations that overlapped with in the new data (32+53-1+11+48-21=122). The SNP set contained 490,362 SNPs that the four datasets shared (486,592 autosomal, 3,757 X-chromosomal, 13 XY). Genotypes given for opposite strands were converted to match HapMap genotypes.

The population counts account for several groupings of samples to form individual populations. In particular, one Djerban Jewish sample was included with the Tunisian Jews. Newly sampled Palestinians were merged with HGDP Palestinians. Newly sampled Jewish individuals and samples from Behar et al. were merged by population. Uzbek Jewish samples contributed by E. Heyer were merged with Uzbek Jewish samples of Behar et al. Two non-Jewish Spanish samples from Behar et al. were merged to one population. Three Ethiopian non-Jewish samples from Behar et al. were treated as one population. One Portuguese Jewish sample from Behar et al. was merged with two new Spanish Jewish samples, and we used the label Iberian Jews for this group. We excluded eight of the 122 populations that were not from Africa, Asia, or Europe (Mexican Americans, Native Americans, Oceanians), leaving 2,789 individuals and 114 distinct for population structure analyses (**Table S1**). **Figure S7** summarizes the merging.

Regional Classifications. Some ambiguity exists regarding the regions with which some populations should be associated, particularly in the Caucasus region. Because Jewish populations from Georgia and Azerbaijan are classified as Mizrahi, we classified nearby non-Jewish populations with the Middle East. Regional classifications for Jewish and non-Jewish populations appear in **Table S1**.

Geographic Coordinates. We assembled a set of geographic coordinates for the populations in the study, taking HGDP coordinates from Rosenberg (2011) based on Cann et al. (2002), and HapMap Luhya and Maasai coordinates from HapMap.

For many populations, we assigned approximate coordinates, in many cases using the same values used by Novembre et al. (2008), MapQuest, or coordinates based on weighted averages, noting that the samples were aggregated from multiple locations. **Table S11** provides the coordinates. The coordinates were used only for visualizing the samples/

Pruned Set of Markers. A pruned set of 5,233 markers was used in *STRUCTURE* analysis, chosen such that adjacent markers were separated by at least 500kb. Beginning from the sorted list of 544,749 autosomal SNPs employed in *RELPAIR* analysis, the first marker on each chromosome was included in the subset. Additional SNPs on a chromosome were chosen from the sorted SNP list to be separated by at least 500kb from the previously selected SNP on the chromosome, ensuring that all 5,233 SNPs selected were in the set of 486,592 used for all population-genetic analyses.

Supplementary References

Boehnke M, Cox NJ (1997) Accurate inference of relationships in sib-pair linkage studies. *Am J Hum Genet* 61: 423-429.

Cann HM, et al. (2002) A human genome diversity cell line panel. *Science* 202: 262-262.

Epstein M, Duren WL, Boehnke M (2000) Improved inference of relationship for pairs of individuals. *Am J Hum Genet* 67: 1219-1231.

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

Jakobsson M, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998-1003.

Matise TC, et al. (2007) A second-generation combined linkage–physical map of the human genome. *Genome Res* 17: 1783-1786.

Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456: 98-101.

Pemberton TJ, Wang C, Li JZ, Rosenberg NA (2010) Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *Am J Hum Genet* 87: 457-464.

Pemberton TJ, et al. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91: 275-292.

Rosenberg NA (2004) *Distruct*: a program for the graphical display of population structure. *Mol Ecol Notes* 4: 137-138.

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet* 70: 841-847.

Rosenberg NA (2011) A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum Biol* 83: 659-684.

Wang S, et al. (2007) Genetic variation and population structure in Native Americans. *PLoS Genet* 3: 2049-2067.

Weir BS (1996) *Genetic Data Analysis II*. Sunderland, MA: Sinauer.

Table S1. Samples used in analysis of population structure. For each population, number of samples in the final combined dataset, dataset of origin, and regional classification are indicated. All samples were from new genotyping, Behar et al. (2010), or the the HGDP or HapMap datasets (Li et al. 2008; Pemberton et al. 2010). Jewish populations are identified as Ashkenazi, Mizrahi, North African, or Sephardi; four Jewish populations are not treated as belonging to any of these groups: EthiopianJ, IndianJ (Cochin), IndianJ (Mumbai), and YemeniteJ. A collection of 145 HGDP and HapMap samples from eight Mexican American, Native American, and Oceanian populations (7 Colombian, 13 Karitiana, 21 Maya, 11 Melanesian, 54 Mexican American MXL, 17 Papuan, 14 Pima, 8 Surui) was included in the final merged data set, but was not used in any analysis. The table thus contains 114 populations: 32 newly sampled populations, 48 Behar et al. including 21 that overlap with the new samples, and 56 HGDP/HapMap including 1 that overlaps with the new samples.

	Number of samples				
Population	**New**	**Behar et al.**	**HGDP/ HapMap**	**Total**	**Regional classification**
Adygei	0	0	17	17	Europe
African American_ASW	0	0	52	52	Africa
AlgerianJ	2	0	0	2	Jewish (North African)
Armenian	0	19	0	19	Middle East
AzerbaijanJ	0	8	0	8	Jewish (Mizrahi)
Balochi	0	0	24	24	Central/South Asia
Bantu (Kenya)	0	0	11	11	Africa
Bantu (S. Africa)	0	0	8	8	Africa
Basque	0	0	24	24	Europe
Bedouin	0	0	45	45	Middle East
Belorussian	0	9	0	9	Europe
BelorussianJ	1	2	0	3	Jewish (Ashkenazi)
Biaka Pygmy	0	0	22	22	Africa
Brahui	0	0	25	25	Central/South Asia
BulgarianJ	19	7	0	26	Jewish (Sephardi)
Burusho	0	0	25	25	Central/South Asia
Cambodian	0	0	10	10	East Asia
Chinese_CHD	0	0	106	106	East Asia
Chuvash	0	17	0	17	Europe
Cypriot	0	12	0	12	Middle East
CzechJ	10	0	0	10	Jewish (Ashkenazi)
Dai	0	0	10	10	East Asia
Daur	0	0	9	9	East Asia
Druze	0	0	42	42	Middle East
DutchJ	1	3	0	4	Jewish (Ashkenazi)
Egyptian	0	12	0	12	Middle East
EgyptianJ	2	0	0	2	Jewish (North African)
Estonian	0	2	0	2	Europe
Ethiopian	0	19	0	19	Africa
EthiopianJ	19	3	0	22	Jewish (Other)
French	0	0	28	28	Europe
FrenchJ	0	1	0	1	Jewish (Ashkenazi)
Georgian	0	20	0	20	Middle East
GeorgianJ	21	4	0	25	Jewish (Mizrahi)
GermanJ	12	2	0	14	Jewish (Ashkenazi)

Gujarati_GIH	0	0	97	97	Central/South Asia
Han	0	0	34	34	East Asia
Han (N. China)	0	0	10	10	East Asia
Han_CHB	0	0	137	137	East Asia
Hazara	0	0	22	22	Central/South Asia
Hezhen	0	0	9	9	East Asia
Hungarian	0	19	0	19	Europe
HungarianJ	21	2	0	23	Jewish (Ashkenazi)
IberianJ	2	1	0	3	Jewish (Sephardi)
Indian (S. India)	0	18	0	18	Central/South Asia
IndianJ (Cochin)	14	3	0	17	Jewish (Other)
IndianJ (Mumbai)	1	4	0	5	Jewish (Other)
Iranian	0	19	0	19	Middle East
IranianJ	20	3	0	23	Jewish (Mizrahi)
IraqiJ	25	8	0	33	Jewish (Mizrahi)
Italian	0	0	12	12	Europe
ItalianJ	1	0	0	1	Jewish (Ashkenazi)
Japanese	0	0	28	28	East Asia
Japanese_JPT	0	0	113	113	East Asia
Jordanian	0	20	0	20	Middle East
Kalash	0	0	23	23	Central/South Asia
Karaite	5	0	0	5	Middle East
KurdishJ	8	0	0	8	Jewish (Mizrahi)
Lahu	0	0	8	8	East Asia
LatvianJ	5	2	0	7	Jewish (Ashkenazi)
Lebanese	0	7	0	7	Middle East
Lezgin	0	18	0	18	Europe
LibyanJ	19	0	0	19	Jewish (North African)
Lithuanian	0	10	0	10	Europe
LithuanianJ	16	1	0	17	Jewish (Ashkenazi)
Luhya_LWK	0	0	99	99	Africa
Maasai_MKK	0	0	105	105	Africa
Makrani	0	0	25	25	Central/South Asia
Mandenka	0	0	22	22	Africa
Mbuti Pygmy	0	0	13	13	Africa
Miao	0	0	10	10	East Asia
Mongolian	0	9	0	9	East Asia
Mongolian (HGDP)	0	0	10	10	East Asia
Moroccan	0	10	0	10	Middle East
MoroccanJ	24	15	0	39	Jewish (North Africa)
Mozabite	0	0	27	27	Middle East
Naxi	0	0	8	8	East Asia
Northern European_CEU	0	0	112	112	Europe
Orcadian	0	0	15	15	Europe
Oroqen	0	0	9	9	East Asia
Palestinian	6	0	46	52	Middle East
Pathan	0	0	22	22	Central/South Asia
PolishJ	20	3	0	23	Jewish (Ashkenazi)
Romanian	0	16	0	16	Europe
RomanianJ	17	3	0	20	Jewish (Ashkenazi)
Russian	0	0	25	25	Europe
RussianJ	26	1	0	27	Jewish (Ashkenazi)

Samaritan	0	2	0	2	Middle East
San	0	0	5	5	Africa
Sardinian	0	0	28	28	Europe
Saudi Arabian	0	20	0	20	Middle East
She	0	0	10	10	East Asia
Sindhi	0	0	24	24	Central/South Asia
Spanish	0	12	0	12	Europe
Syrian	0	16	0	16	Middle East
Tajik	16	0	0	16	Central/South Asia
Toscani_TSI	0	0	102	102	Europe
Tu	0	0	10	10	East Asia
Tujia	0	0	10	10	East Asia
TunisianJ	29	0	0	29	Jewish (North African)
Turkish	0	19	0	19	Middle East
TurkishJ	14	10	0	24	Jewish (Sephardi)
Tuscan	0	0	7	7	Europe
UkrainianJ	18	0	0	18	Jewish (Ashkenazi)
Uygur	0	0	10	10	Central/South Asia
Uzbek	0	15	0	15	Central/South Asia
UzbekJ	9	2	0	11	Jewish (Mizrahi)
Xibo	0	0	9	9	East Asia
Yakut	0	0	25	25	East Asia
Yemenite	0	8	0	8	Middle East
YemeniteJ	26	14	0	40	Jewish (Other)
Yi	0	0	10	10	East Asia
Yoruba	0	0	21	21	Africa
Yoruba_YRI	0	0	140	140	Africa
Total	**429**	**450**	**1910**	**2789**	

Table S2. Sample sets. Twelve sample sets used in population-genetic data analyses.

Set number	Description	Number of individuals	Number of populations	Number of Jewish individuals	Number of Jewish populations	Figures using the set
1	Jewish + Africa + Asia + Europe	2789	114	504	31	1A
2	Jewish + C/S Asia + Europe + Middle East – Ethiopian Jews	1656	79	482	30	1B
3	Jewish + Europe + Middle East – Ethiopian Jews – Indian Jews (Cochin) – Indian Jews (Mumbai)	1288	64	460	28	1C, 2A, 2B
4	Jewish – Ethiopian Jews – Indian Jews (Cochin) – Indian Jews (Mumbai) – Yemenite Jews	420	27	420	27	3A, 3B
5	Ashkenazi Jewish + European non-Jewish	632	31	159	13	4A
6	Mizrahi Jewish + Middle Eastern non-Jewish	179	10	104	6	4B
7	North African Jewish + North African non-Jewish	140	8	91	5	4C
8	Sephardi Jewish + Mediterranean non-Jewish	131	9	53	3	4D
9	Ashkenazi Jewish	159	13	159	13	4E, 5A
10	Mizrahi Jewish	104	6	104	6	4F, 5C
11	North African Jewish	91	5	91	5	4G, 5D
12	Sephardi Jewish	53	3	53	3	4H, 5B

Table S3. Mean and standard deviation across loci of expected heterozygosity, for Middle Eastern, European, and Jewish populations. The table contains 67 populations: 64 populations from sample set 2 (**Table S2**), plus Ethiopian Jews, Indian Jews (Cochin), and Indian Jews (Mumbai).

Population	Regional classification	Number of individuals	Mean expected heterozygosity across loci	Standard deviation of expected heterozygosity across loci
Adygei	Europe	17	0.3254	0.1619
AlgerianJ	Jewish (North African)	2	0.3248	0.2856
Armenian	Middle East	19	0.3257	0.1606
AzerbaijaniJ	Jewish (Mizrahi)	8	0.3189	0.1827
Basque	Europe	24	0.3186	0.1620
Bedouin	Middle East	45	0.3266	0.1524
Belorussian	Europe	9	0.3225	0.1756
BelorussianJ	Jewish (Ashkenazi)	3	0.3251	0.2318
BulgarianJ	Jewish (Sephardi)	26	0.3260	0.1562
Chuvash	Europe	17	0.3259	0.1618
Cypriot	Middle East	12	0.3248	0.1681
CzechJ	Jewish (Ashkenazi)	10	0.3244	0.1724
Druze	Middle East	42	0.3231	0.1556
DutchJ	Jewish (Ashkenazi)	4	0.3255	0.2094
Egyptian	Middle East	12	0.3350	0.1612
EgyptianJ	Jewish (North African)	2	0.3125	0.2815
Estonian	Europe	2	0.3242	0.2803
EthiopianJ	Jewish (Other)	22	0.3335	0.1537
French	Europe	28	0.3243	0.1559
FrenchJ	Jewish (Ashkenazi)	1	0.3245	0.4682
Georgian	Middle East	20	0.3236	0.1617
GeorgianJ	Jewish (Mizrahi)	25	0.3223	0.1601
GermanJ	Jewish (Ashkenazi)	14	0.3233	0.1661
Hungarian	Europe	19	0.3247	0.1598
HungarianJ	Jewish (Ashkenazi)	23	0.3250	0.1583
IberianJ	Jewish (Sephardi)	3	0.3252	0.2321
IndianJ (Cochin)	Jewish (Other)	17	0.3181	0.1688
IndianJ (Mumbai)	Jewish (Other)	5	0.3145	0.2026
Iranian	Middle East	19	0.3301	0.1570
IranianJ	Jewish (Mizrahi)	23	0.3224	0.1612
IraqiJ	Jewish (Mizrahi)	33	0.3233	0.1572
Italian	Europe	12	0.3233	0.1683
ItalianJ	Jewish (Ashkenazi)	1	0.3256	0.4686
Jordanian	Middle East	20	0.3319	0.1548
Karaite	Middle East	5	0.3096	0.2048
KurdishJ	Jewish (Mizrahi)	8	0.3249	0.1792
LatvianJ	Jewish (Ashkenazi)	7	0.3257	0.1820

Lebanese	Middle East	7	0.3286	0.1813
Lezgin	Europe	18	0.3249	0.1615
LibyanJ	Jewish (North African)	19	0.3225	0.1633
Lithuanian	Europe	10	0.3202	0.1743
LithuanianJ	Jewish (Ashkenazi)	17	0.3245	0.1625
Moroccan	Middle East	10	0.3351	0.1654
MoroccanJ	Jewish (North African)	39	0.3258	0.1537
Mozabite	Middle East	27	0.3284	0.1550
Northern European_CEU	Europe	112	0.3246	0.1487
Orcadian	Europe	15	0.3218	0.1657
Palestinian	Middle East	52	0.3289	0.1497
PolishJ	Jewish (Ashkenazi)	23	0.3247	0.1587
Romanian	Europe	16	0.3262	0.1614
RomanianJ	Jewish (Ashkenazi)	20	0.3249	0.1598
Russian	Europe	25	0.3246	0.1571
RussianJ	Jewish (Ashkenazi)	27	0.3258	0.1563
Samaritan	Middle East	2	0.2822	0.2819
Sardinian	Europe	28	0.3179	0.1611
Saudi Arabian	Middle East	20	0.3262	0.1594
Spanish	Europe	12	0.3248	0.1671
Syrian	Middle East	16	0.3290	0.1600
Toscani_TSI	Europe	102	0.3248	0.1497
TunisianJ	Jewish (North African)	29	0.3250	0.1563
Turkish	Middle East	19	0.3281	0.1581
TurkishJ	Jewish (Sephardi)	24	0.3264	0.1567
Tuscan	Europe	7	0.3248	0.1824
UkrainianJ	Jewish (Ashkenazi)	18	0.3256	0.1606
UzbekJ	Jewish (Mizrahi)	11	0.3225	0.1724
Yemenite	Middle East	8	0.3372	0.1704
YemeniteJ	Jewish (Other)	40	0.3226	0.1563

Table S4. **Mean across loci of expected heterozygosity in several sets of populations.**

Group	Number of individuals	Number of populations	Mean heterozygosity across populations	Heterozygosity for the pooled set of individuals
Europe	473	18	0.3235	0.3259
Middle East	355	18	0.3252	0.3311
Jewish	504	31	0.3237	0.3291
Ashkenazi Jewish	168	13	0.3249	0.3251
Mizrahi Jewish	108	6	0.3224	0.3252
North African Jewish	91	5	0.3221	0.3266
Sephardi Jewish	53	3	0.3259	0.3265

Table S5. Number of replicates placed by *CLUMPAK* in the major mode. The total number of *STRUCTURE* replicates in each analysis was 20 for each *K*.

			K						
Set number	**Description**	**Figure**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
3	Jewish + Europe + Middle East – Ethiopian Jews – Indian Jews (Cochin) – Indian Jews (Mumbai)	2A	20	20	18	17	11	13	14
4	Jewish – Ethiopian Jews – Indian Jews (Cochin) – Indian Jews (Mumbai) – Yemenite Jews	3B	20	20	18	15	9	14	10
9	Ashkenazi Jewish	5A	13	17	18	18	18	-	-
10	Mizrahi Jewish	5C	14	9	11	16	14	-	-
11	North African Jewish	5D	20	15	18	13	17	-	-
12	Sephardi Jewish	5B	20	14	18	18	16	-	-

Table S6. Outliers. Thirteen individuals removed in MDS and *STRUCTURE* analyses of subgroups of Jewish populations.

Identification number	Sample source	Population
5667	NLGIP	CzechJ
1972	NLGIP	GeorgianJ
1961	NLGIP	GeorgianJ
GeorgianJew125	Behar et al. (2010)	GeorgianJ
1165	NLGIP	IraqiJ
6256	NLGIP	LatvianJ
1583	NLGIP	RussianJ
1690	NLGIP	RussianJ
1742	NLGIP	RussianJ
1801	NLGIP	RussianJ
1962	NLGIP	RussianJ
4211	NLGIP	RussianJ
1978	NLGIP	UkrainianJ

Table S7. Individuals new for this study. Individuals with a relationship more distant than second-degree (avuncular, half sib, or grandparent/grandchild) were treated as unrelated. Samples collected for this study were collected at the Barzilai Medical Center in Ashkelon, Israel, with the exception of the Karaite samples which were collected at a Karaite community center in Ashkelon, Israel. The samples from E. Heyer were obtained at Bukhara, Uzbekistan; the population labeled "Tajik" represents Tajik-speaking individuals from Bukhara. For data analyses, the Tunisian Jewish sample from Djerba was included with the Tunisian Jewish population, so that we regard the new sample set as containing 29 Jewish and 3 non-Jewish populations.

Population	NLGIP samples	Sampled by E. Heyer	Newly sampled in Ashkelon, Israel	Total collected	Total passing genotyping center quality control	Number of duplicates removed	Number of relatives removed	Number of individuals removed for other reasons	Number of unrelated individuals in the final dataset
AlgerianJ	0	0	2	2	2				2
BelorussianJ	3	0	1	4	1				1
BulgarianJ	20	0	0	20	20	1			19
CzechJ	10	0	0	10	10				10
DutchJ	2	0	0	2	1				1
EgyptianJ	0	0	2	2	2				2
EthiopianJ	20	0	3	23	21	2			19
GeorgianJ	20	0	2	22	21				21
GermanJ	12	0	0	12	12				12
HungarianJ	26	0	2	28	23	2			21
IndianJ (Cochin)	20	0	0	20	19		5		14
IndianJ (Mumbai)	0	0	1	1	1				1
IranianJ	20	0	7	27	20				20
IraqiJ	20	0	9	29	25				25
ItalianJ	1	0	0	1	1				1
Karaite	0	0	8	8	5				5
KurdishJ	9	0	4	13	9	1			8
LatvianJ	6	0	0	6	5				5
LibyanJ	20	0	5	25	20		1		19
LithuanianJ	19	0	0	19	16				16
MoroccanJ	15	0	12	27	24				24
Palestinian	16	0	5	21	16		1		15
PolishJ	20	0	4	24	21		1		20
RomanianJ	17	0	0	17	17				17

RussianJ	27	0	0	27	26				26
SpanishJ	0	0	4	4	2				2
Tajik	0	20	0	20	16				16
TunisianJ	24	0	11	35	28				28
TunisianJ (Djerba)	0	0	1	1	1				1
TurkishJ	20	0	1	21	14				14
UkrainianJ	20	0	0	20	18				18
UzbekJ	0	17	0	17	9				9
YemeniteJ	20	0	7	27	26				26
UnknownJ	0	0	0	0	4	1		3	0
Total	**407**	**37**	**91**	**535**	**456**	**7**	**8**	**3**	**438**

Table S8. Pairs of duplicate samples among the samples collected for this study in combination with NLGIP samples. Each of these pairs contained two samples among those obtained from NLGIP. For each pair, the sample with the larger rate of missing data was omitted.

Duplicate pair number	IDs of duplicate samples	Populations	ID of omitted sample
1	4752, 4783	Hungarian Jewish & Unknown	4752
2	5157, 5285	Bulgarian Jewish	5157
3	1599, 4657	Ethiopian Jewish	4657
4	5712, 6022	Hungarian Jewish	6022
5	6191, 6245	Hungarian Jewish	6245
6	1580, 4584	Kurdish Jewish	1580
7	1822, 1822	Ethiopian Jewish	1822

Table S9. Samples removed due to relatedness among the samples collected for this study in combination with NLGIP samples. Relatedness was inferred using identity-in-state allele sharing and *RELPAIR* (see **Figure S3**).

Population	Number of pairs of close relatives	IDs of omitted samples	Total number of individuals after removal of relatives
Indian Jewish (Cochin)	7	40041, 40171, 40302, 40322, 40451	14
Libyan Jewish	1	1601	19
Palestinian	1	4926	15
Polish Jewish	1	1106	20

Table S10. Samples incorporated in this study from Behar et al. (2010). Individuals with a relationship more distant than second-degree (avuncular, half sib, or grandparent/grandchild) were treated as unrelated. Treating the three Ethiopian non-Jewish samples as one population and the two Spanish non-Jewish samples as one population, the table contains 48 populations.

Population	Behar et al. (2010) samples	Number of duplicates removed	Number of relatives removed	Number of individuals removed for other reasons	Number of unrelated individuals in Behar et al. (2010)	Additional duplicates and relatives due to combining Jewish datasets	Number of individuals in the "final combined set"
Armenian	19				19		19
AzerbaijanJ	8				8		8
Belorussian	10			1	9		9
BelorussianJ	2				2		2
BulgarianJ	8				8	1 (duplicate)	7
Chuvash	17				17		17
Cypriot	12				12		12
DutchJ	3				3		3
Egyptian	12				12		12
Estonian	2				2		2
Ethiopian (Amhara)	7				7		7
Ethiopian (Oromo)	7				7		7
Ethiopian (Tigray)	5				5		5
EthiopianJ	13	1			12	9 (duplicates)	3
FrenchJ	1				1		1
Georgian	20				20		20
GeorgianJ	4				4		4
GermanJ	2				2		2
Hungarian	20		1		19		19
HungarianJ	3				3	1 (duplicate)	2
Indian (S. India)	19		1		18		18
IndianJ (Cochin)	4				4	1 (relative)	3
IndianJ (Mumbai)	4				4		4
Iranian	20		1		19		19
IranianJ	4				4	1 (duplicate)	3
IraqiJ	11		1		10	2 (duplicates)	8
Jordanian	20				20		20

LatvianJ	2				2		2
Lebanese	8			1	7		7
Lezgin	18				18		18
Lithuanian	10				10		10
LithuanianJ	1				1		1
Mongolian	9				9		9
Moroccan	10				10		10
MoroccanJ	17	1		1	15		15
PolishJ	3				3		3
PortugueseJ	1				1		1
Romanian	16				16		16
RomanianJ	3				3		3
RussianJ	1				1		1
Samaritan	3		1		2		2
Saudi Arabian	20				20		20
Spanish (Andalusia)	6				6		6
Spanish (Catalonia)	6				6		6
Syrian	16				16		16
Turkish	19				19		19
TurkishJ	10				10		10
Uzbek	15				15		15
UzbekJ	2				2		2
Yemenite	10	1	1		8		8
YemeniteJ	15				15	1 (duplicate)	14
Total	**478**	**3**	**6**	**3**	**466**	**16**	**450**

Table S11. Geographic coordinates. Coordinates were used to generate a map of population locations (**Figure S1**). Corresponding Jewish and non-Jewish populations are sometimes but not always assigned the same geographic coordinates. We took coordinates from HGDP and HapMap sources for populations from those datasets; for other populations, we assigned approximate coordinates, from Novembre et al. (2008), from MapQuest, or by methods briefly described in the table. The coordinates do not necessarily accord with sampling locations and are used only to assist in visualzing the sampled data set rather than as a formal basis for data analysis. The table contains 112 populations that appear on the map in **Figure 1**, considering 3 Ethiopian non-Jewish samples as separate groups. Four HapMap populations used in the data analysis (**Table S1**) do not appear in **Figure 1** and are not included in the table (African American_ASW, Chinese_CHD, Northern European_CEU, and Gujarati_GIH).

Population	Latitude	Longitude	Source
Adygei	44	39	HGDP (Rosenberg 2011)
AlgerianJ	31	2.2	MapQuest
Armenian	40.3	44.9	MapQuest
AzerbaijaniJ	40.2	49	MapQuest
Balochi	30.5	66.5	HGDP (Rosenberg 2011)
Bantu (Kenya)	-3	37	HGDP (Rosenberg 2011)
Bantu (S. Africa)	-25.6	24.3	HGDP (Rosenberg 2011)
Basque	43	0	HGDP (Rosenberg 2011)
Bedouin	31	35	HGDP (Rosenberg 2011)
Belorussian	53.5	28.1	MapQuest
BelorussianJ	53.5	28.1	MapQuest
Biaka Pygmy	4	17	HGDP (Rosenberg 2011)
Brahui	30.5	66.5	HGDP (Rosenberg 2011)
BulgarianJ	42.8	25.2	Novembre et al. (2008)
Burusho	36.5	74	HGDP (Rosenberg 2011)
Cambodian	12	105	HGDP (Rosenberg 2011)
Chuvash	55.5	47	Republic of Chuvashia (approximate)
Cypriot	35.1	33.2	Novembre et al. (2008)
CzechJ	49.7	15.4	Novembre et al. (2008)
Dai	21	100	HGDP (Rosenberg 2011)
Daur	48.5	124	HGDP (Rosenberg 2011)
Druze	32	35	HGDP (Rosenberg 2011)
DutchJ	52.3	5.67	Novembre et al. (2008)
Egyptian	26.3	29.1	MapQuest
EgyptianJ	26.3	29.1	MapQuest
Estonian	58.7	25.8	MapQuest
Ethiopian (Amhara)	11.6	37.4	MapQuest (Bahir Dar, Amhara region)
Ethiopian (Oromo)	9	38.8	MapQuest (Addis Ababa, Oromia region)
Ethiopian (Tigray)	13.5	39.5	MapQuest (Mekelle, Tigray region)
EthiopianJ	8.4	39.1	MapQuest
French	46	2	HGDP (Rosenberg 2011)

FrenchJ	46	2	HGDP (Rosenberg 2011)
Georgian	42.2	43.5	MapQuest
GeorgianJ	42.2	43.5	MapQuest
GermanJ	51.1	10.4	Novembre et al. (2008)
Han	32.3	114	HGDP (Rosenberg 2011)
Han (N. China)	32.3	114	HGDP (Rosenberg 2011)
Han_CHB	39.9	116.4	MapQuest (HapMap 3, Beijing)
Hazara	33.5	70	HGDP (Rosenberg 2011)
Hezhen	47.5	133.5	HGDP (Rosenberg 2011)
Hungarian	47.2	19.4	Novembre et al. (2008)
HungarianJ	47.2	19.4	Novembre et al. (2008)
IberianJ	40.1	-5.1	Weighted mean of Spain and Portugal, Novembre et al. (2008)
Indian (S. India)	12.1	77.3	Weighted mean of Kerala, North Kannadi, and Tamil Nadu regions
IndianJ (Cochin)	10	76.3	MapQuest (Cochin)
IndianJ (Mumbai)	19	72.8	MapQuest (Mumbai)
Iranian	32	54.5	MapQuest
IranianJ	32	54.5	MapQuest
IraqiJ	32.9	43.6	MapQuest
Italian	46	10	HGDP (Rosenberg 2011)
ItalianJ	46	10	HGDP (Rosenberg 2011)
Japanese	38	138	HGDP (Rosenberg 2011)
Japanese_JPT	35.7	139.8	MapQuest (HapMap 3, Tokyo)
Jordanian	32.4	37.8	MapQuest
Kalash	36	71.5	HGDP (Rosenberg 2011)
Karaite	30	31.2	MapQuest (Cairo)
KurdishJ	37.5	43.2	Kurdish-inhabited region (approximate)
Lahu	22	100	HGDP (Rosenberg 2011)
LatvianJ	56.9	24.9	Novembre et al. (2008)
Lebanese	33.9	35.9	MapQuest
Lezgin	41.5	48	Lezgin-inhabited region (approximate)
LibyanJ	28	16.4	MapQuest
Lithuanian	55.3	23.9	MapQuest
LithuanianJ	55.3	23.9	MapQuest
Luhya_LWK	0.6	34.8	MapQuest (HapMap 3, Webuye)
Maasai_MKK	1.8	37.6	MapQuest (HapMap 3, Kinyawa)
Makrani	26	64	HGDP (Rosenberg 2011)
Mandenka	12	-12	HGDP (Rosenberg 2011)
Mbuti Pygmy	1	29	HGDP (Rosenberg 2011)
Miao	28	109	HGDP (Rosenberg 2011)
Mongolian	45	111	HGDP (Rosenberg 2011)

Mongolian (HGDP)	45	111	HGDP (Rosenberg 2011)
Moroccan	30	-8.1	MapQuest
MoroccanJ	30	-8.1	MapQuest
Mozabite	32	3	HGDP (Rosenberg 2011)
Naxi	26	100	HGDP (Rosenberg 2011)
Orcadian	59	-3	HGDP (Rosenberg 2011)
Oroqen	50.4	126.5	HGDP (Rosenberg 2011)
Palestinian	32	35	HGDP (Rosenberg 2011)
Pathan	33.5	70.5	HGDP (Rosenberg 2011)
PolishJ	52.1	19.4	Novembre et al. (2008)
Romanian	45.9	25	Novembre et al. (2008)
RomanianJ	45.9	25	Novembre et al. (2008)
Russian	61	40	HGDP (Rosenberg 2011)
RussianJ	61	40	HGDP (Rosenberg 2011)
Samaritan	32	34.5	MapQuest (Holon)
San	-21	20	HGDP (Rosenberg 2011)
Sardinian	40	9	HGDP (Rosenberg 2011)
Saudi Arabian	24.6	46.7	MapQuest
She	27	119	HGDP (Rosenberg 2011)
Sindhi	25.5	69	HGDP (Rosenberg 2011)
Spanish	40.3	-3.6	Novembre et al. (2008)
Syrian	35.9	38.4	MapQuest
Tajik	39.8	64.4	MapQuest (Bukhara)
Toscani_TSI	43	11	HGDP (Rosenberg 2011)
Tu	36	101	HGDP (Rosenberg 2011)
Tujia	29	109	HGDP (Rosenberg 2011)
TunisianJ	34	9.6	MapQuest
Turkish	39.1	35.4	Novembre et al. (2008)
TurkishJ	39.1	35.4	Novembre et al. (2008)
Tuscan	43	11	HGDP (Rosenberg 2011)
UkrainianJ	49.1	31.4	Novembre et al. (2008)
Uygur	44	81	HGDP (Rosenberg 2011)
Uzbek	42	63.3	MapQuest
UzbekJ	39.8	64.4	MapQuest (Bukhara)
Xibo	43.5	81.5	HGDP (Rosenberg 2011)
Yakut	63	129.5	HGDP (Rosenberg 2011)
Yemenite	14.7	47.6	MapQuest
YemeniteJ	14.7	47.6	MapQuest
Yi	28	103	HGDP (Rosenberg 2011)
Yoruba	8	5	HGDP (Rosenberg 2011)
Yoruba_YRI	7.4	3.9	MapQuest (HapMap, Ibadan)

Figure S1. Populations included in the study. Each population is shown at an approximate geographic location (Table S11). Populations are assigned the same symbol that is used in Figure 1C, or a black circle if not included in that figure.

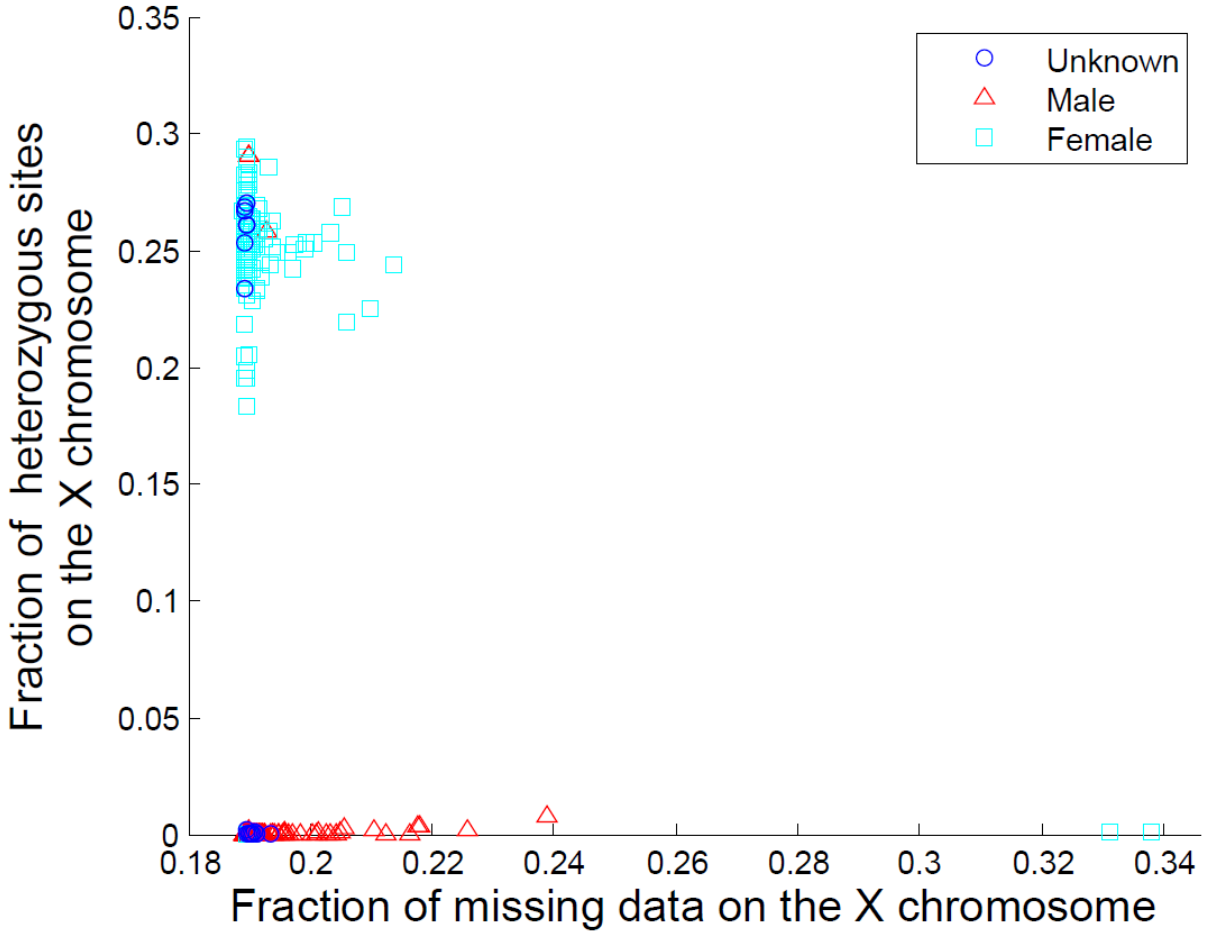
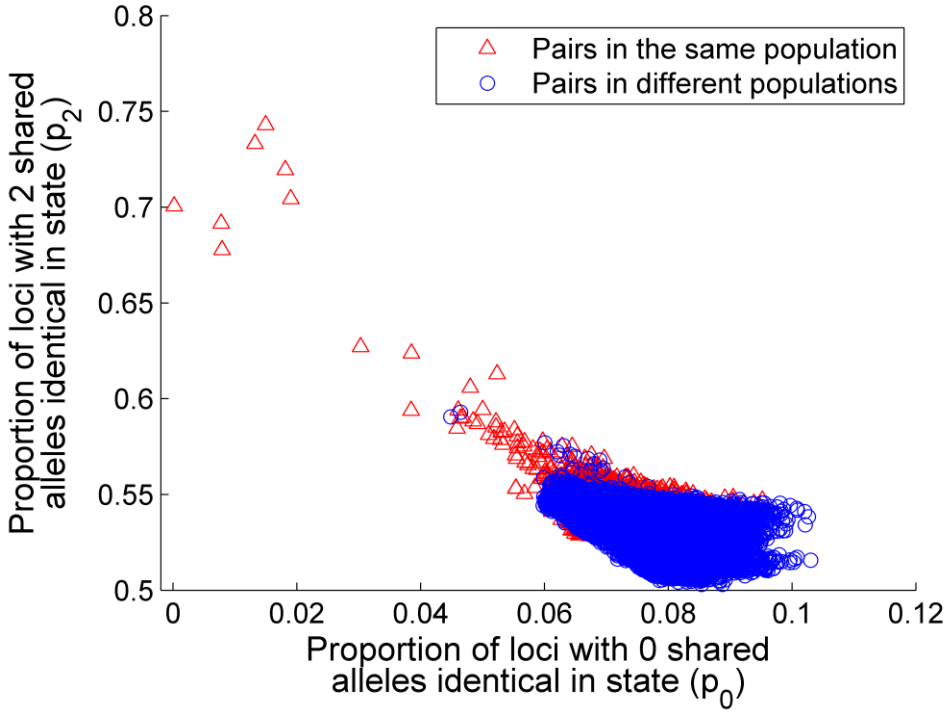


Figure S2. Fraction of heterozygous sites on the X chromosome vs. fraction of sites with missing data on the X chromosome. Samples are sorted by their *a priori* sex assignments. This analysis relied on 456 individuals; sex assignments were corrected for 7 outliers and assigned for 19 samples with no prior information, and no individuals were removed.

A



B

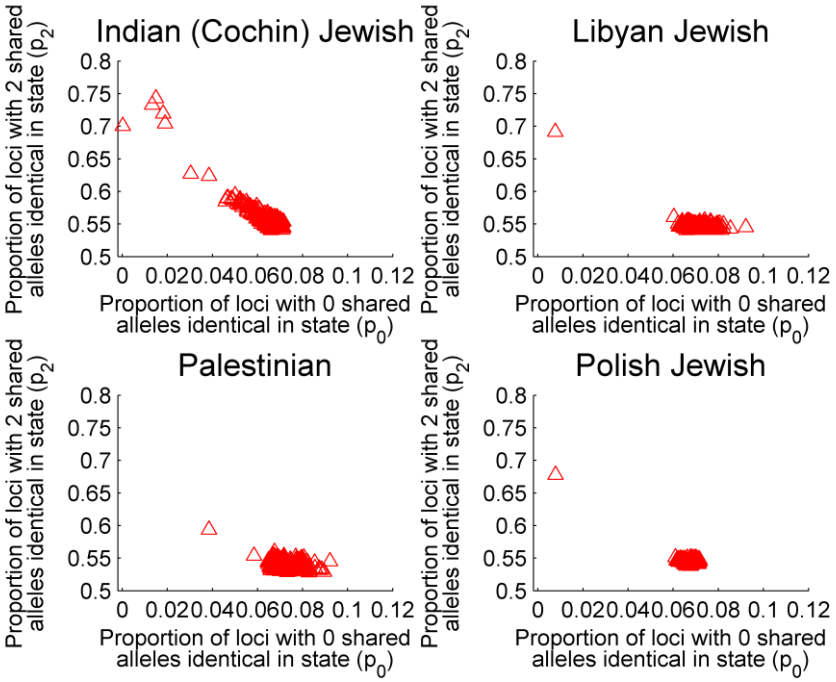


Figure S3. Proportions of loci with 2 and 0 alleles shared identical in state. (A) Pairs of individuals who are not necessarily from the same population. Duplicates were removed for this figure. Relative pairs appear on the far left. The two leftmost blue circles are of a Georgian Jewish and a Russian Jewish pair and a Georgian Jewish and a Ukrainian Jewish pair. (B) Pairs of individuals from specific populations in which relative pairs were detected. This analysis led to the removal of 8 samples (5 Cochin Jewish, 1 Libyan Jewish, 1 Palestinian, 1 Polish Jewish).

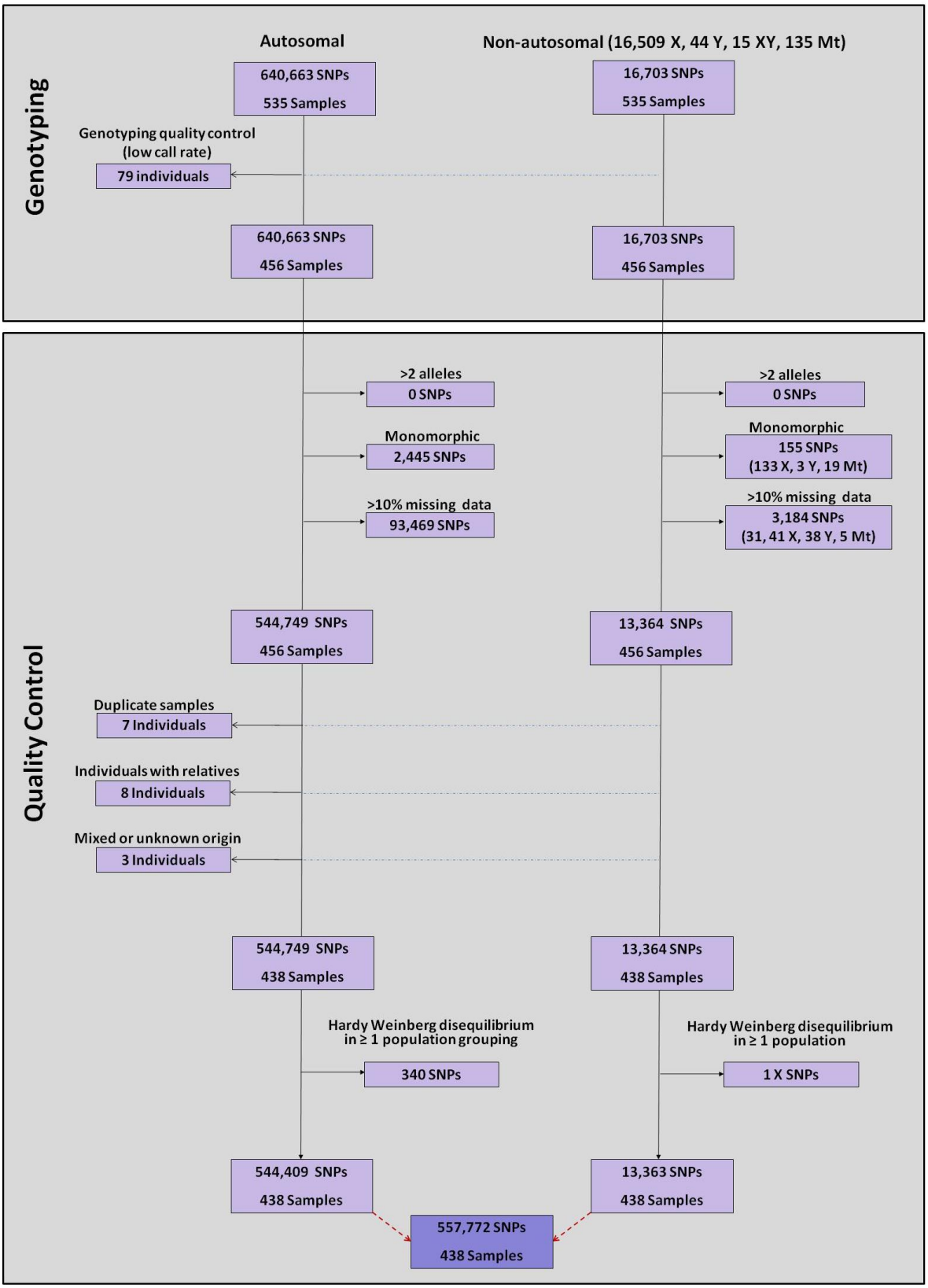


Figure S4. Genotyping and quality control for new SNP genotypes.

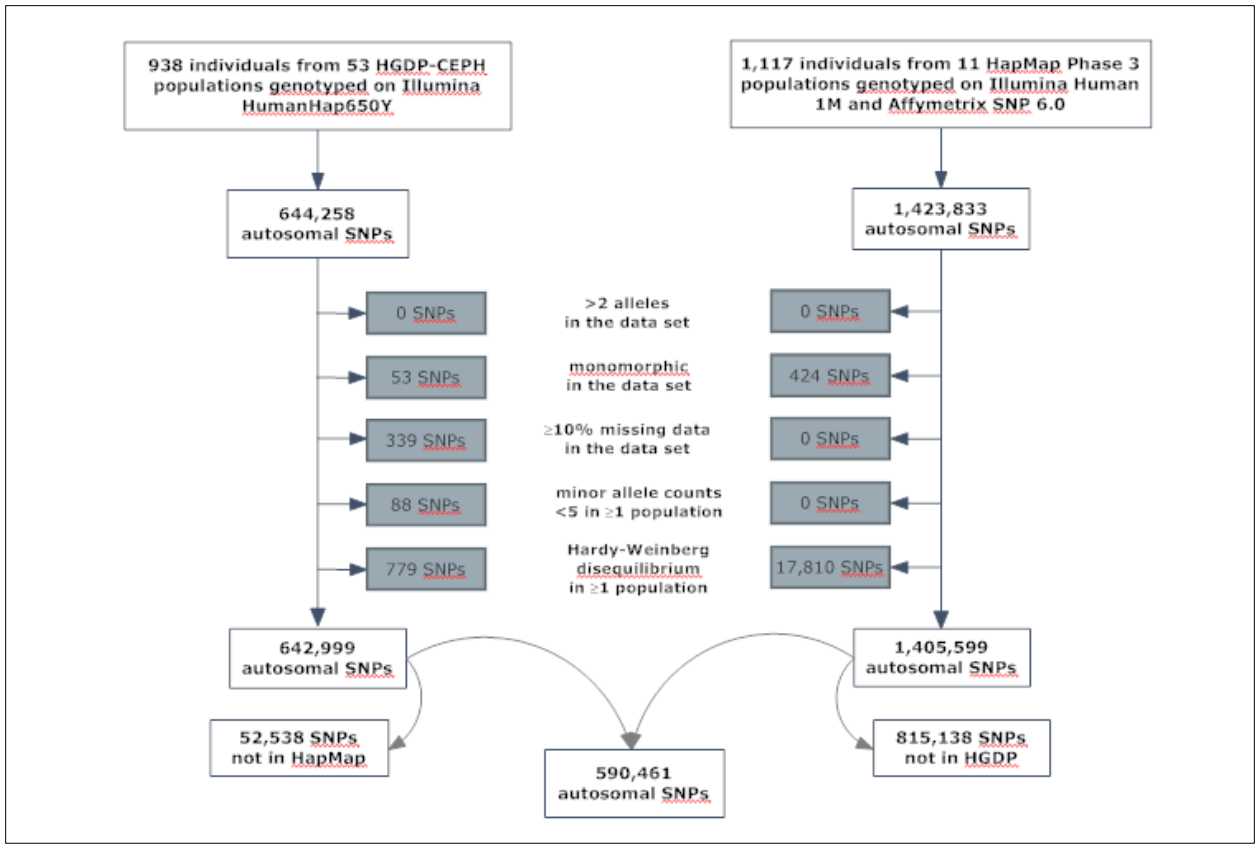


Figure S5. Quality control for HGDP and HapMap genotypes.

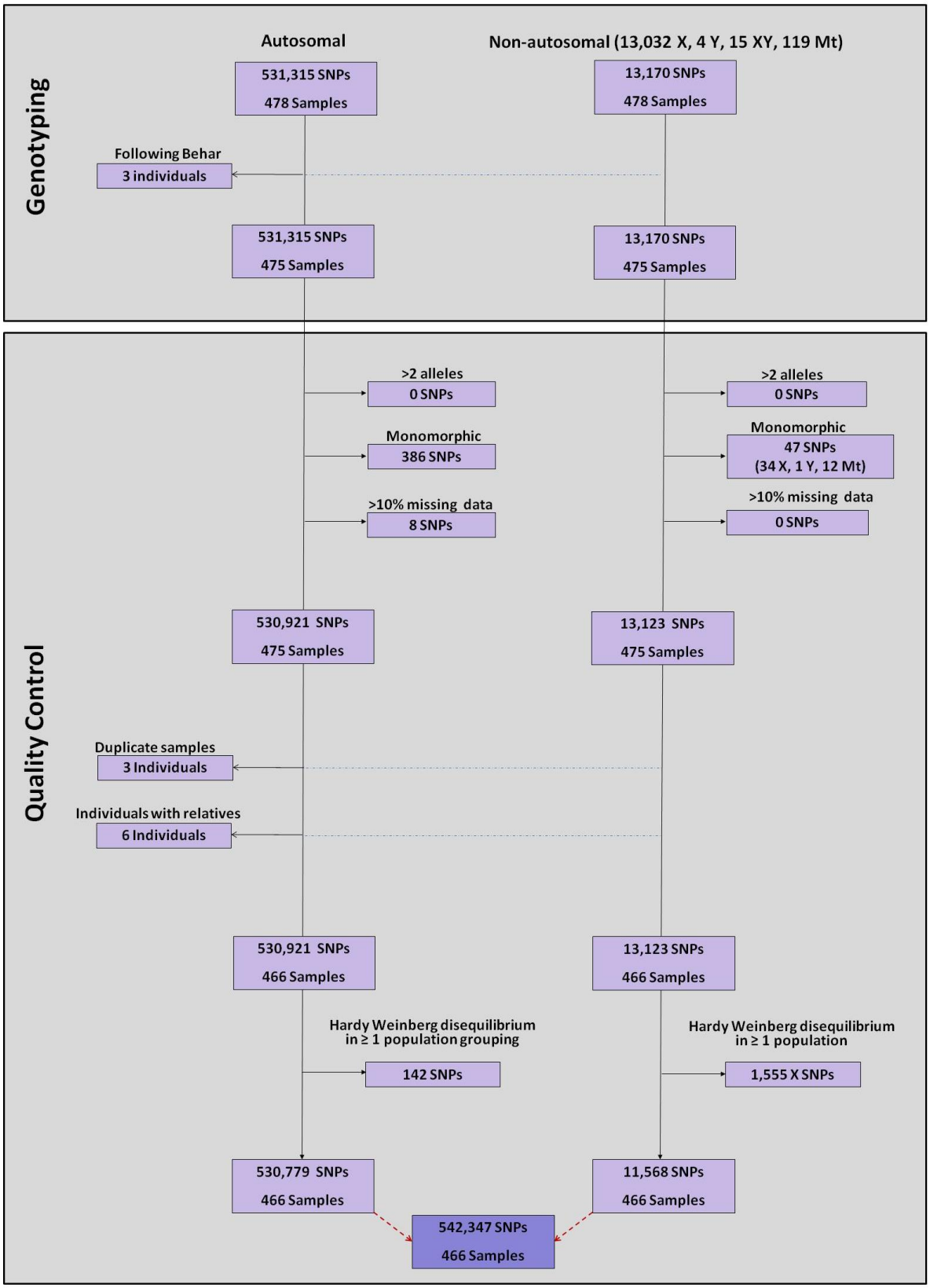


Figure S6. Quality control for SNP genotypes from Behar et al. (2010).

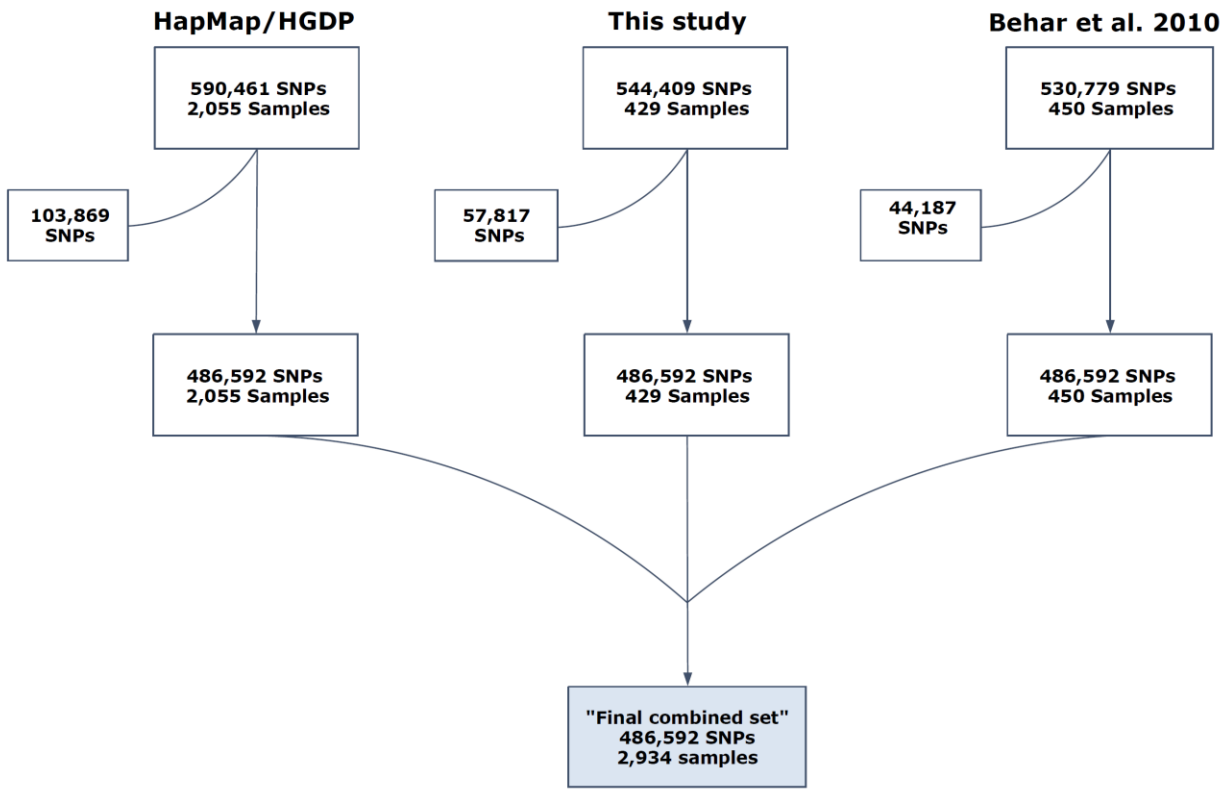


Figure S7. The data merging process that led to the final combined set used in the paper for autosomal markers.

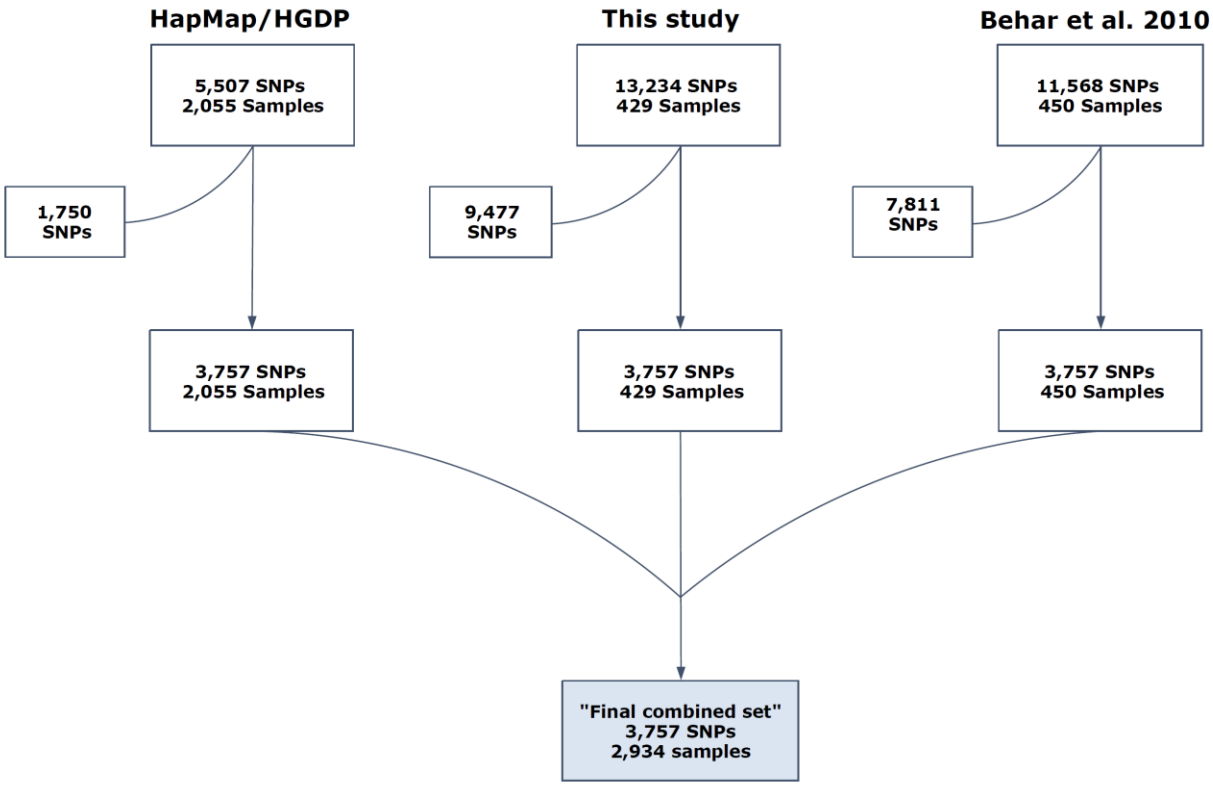


Figure S8. The data merging process that led to the final combined set of X-chromosomal markers.