The American Journal of Human Genetics, Volume *91*

Supplemental Data

Genomic Patterns of Homozygosity in Worldwide Human Populations

Trevor J. Pemberton, Devin Absher, Marcus W. Feldman, Richard M. Myers, Noah A. Rosenberg, and Jun Z. Li

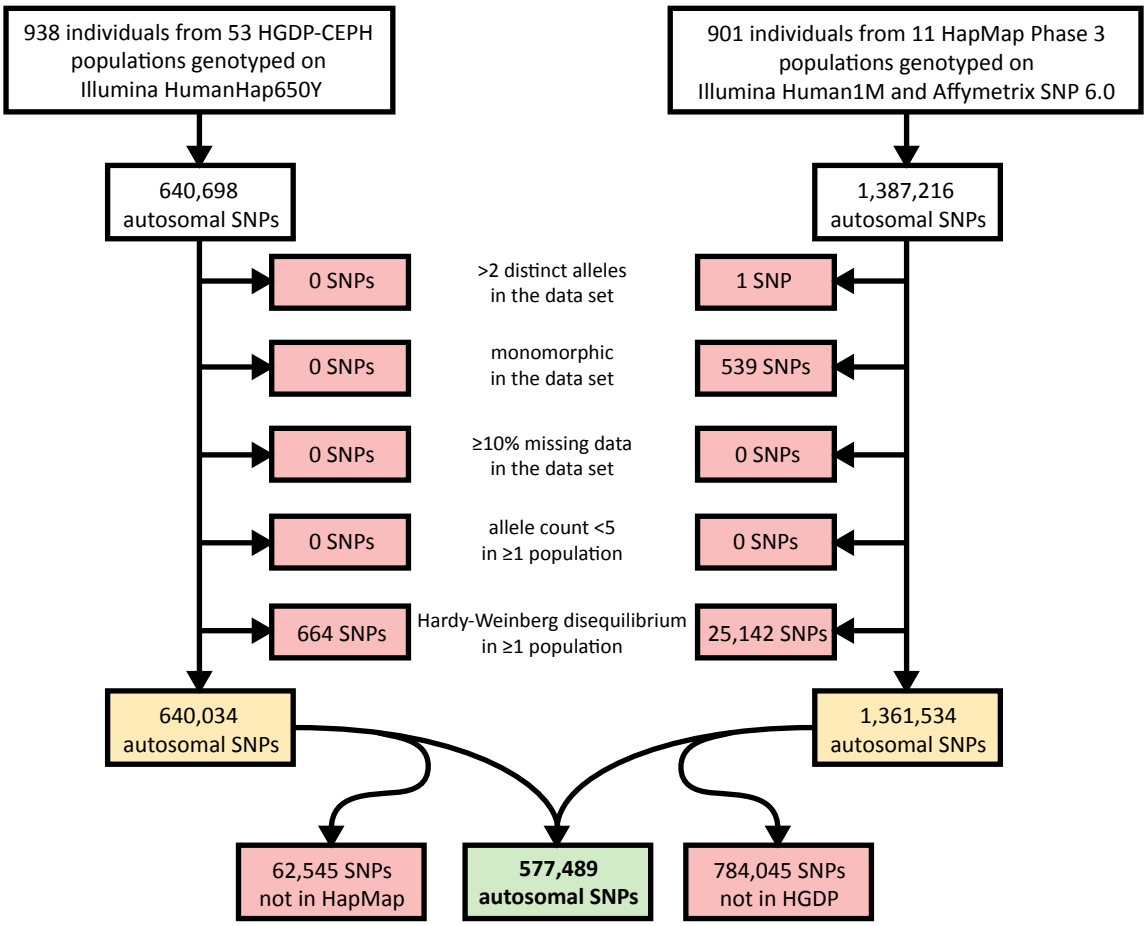

Figure S1. Flow Diagram of the Sequential Data Filtering Steps Used for Preparation of the Combined Data Set

Steps are shown in the order in which they were applied. SNPs that failed in one step were removed and were not considered in subsequent steps. The numbers of SNPs removed at each step are shown in the boxes shaded in red, the final numbers of SNPs in the two data sets are shown in the boxes shaded in orange, and the final number of SNPs in the combined data set is shown in the box shaded in green.

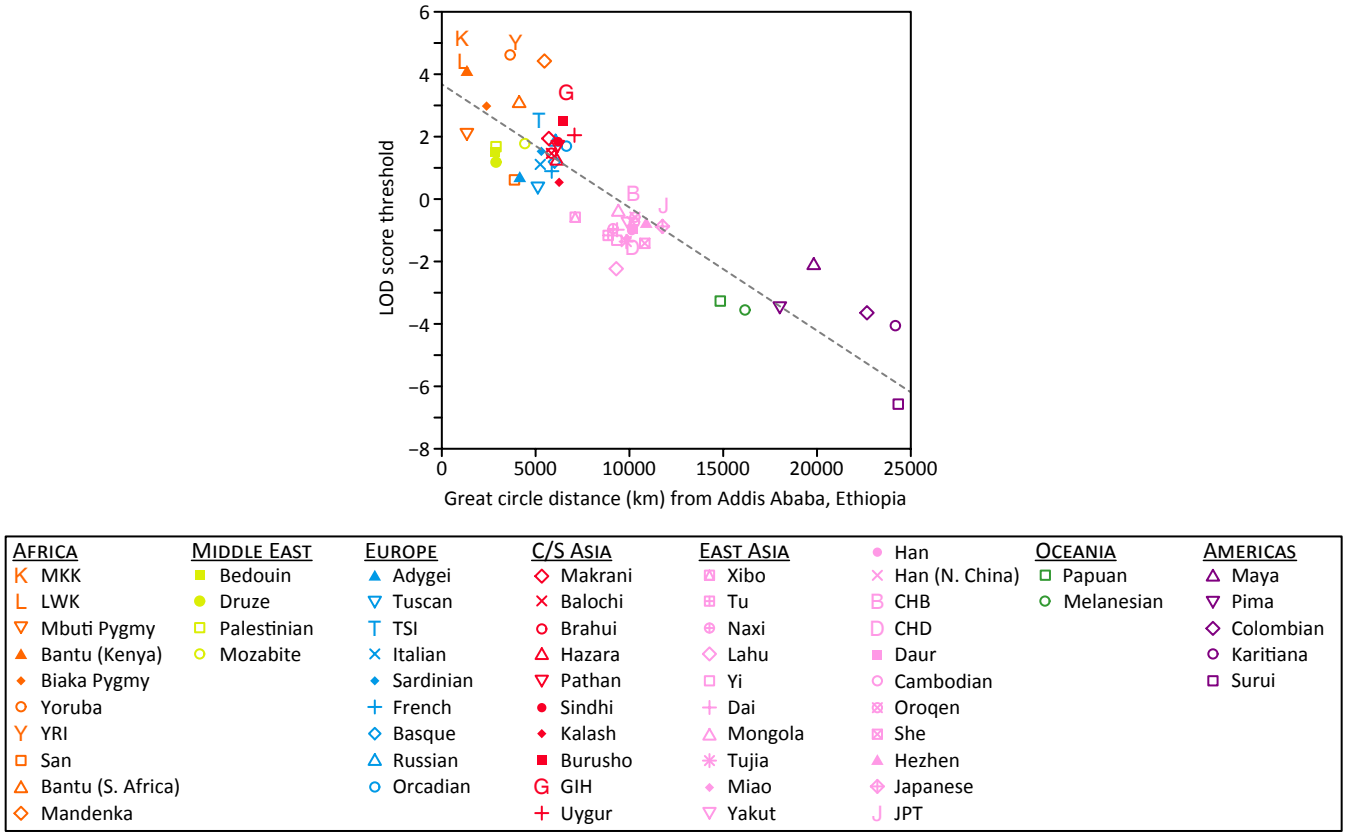

Figure S2. The Decrease of the LOD-Score Threshold with Geographic Distance from Addis Ababa, Ethiopia, an Approximate Location for the Origin of the Out-of-Africa Migration

The coefficient of determination is R^2=0.757. The LOD score thresholds from 61 of the 64 populations are included: the ASW and MXL populations were excluded as they represent admixed populations, and the CEU population was excluded as it represents general Northern European ancestry rather than individuals from a specific location. Populations are colored by their regional affiliation.

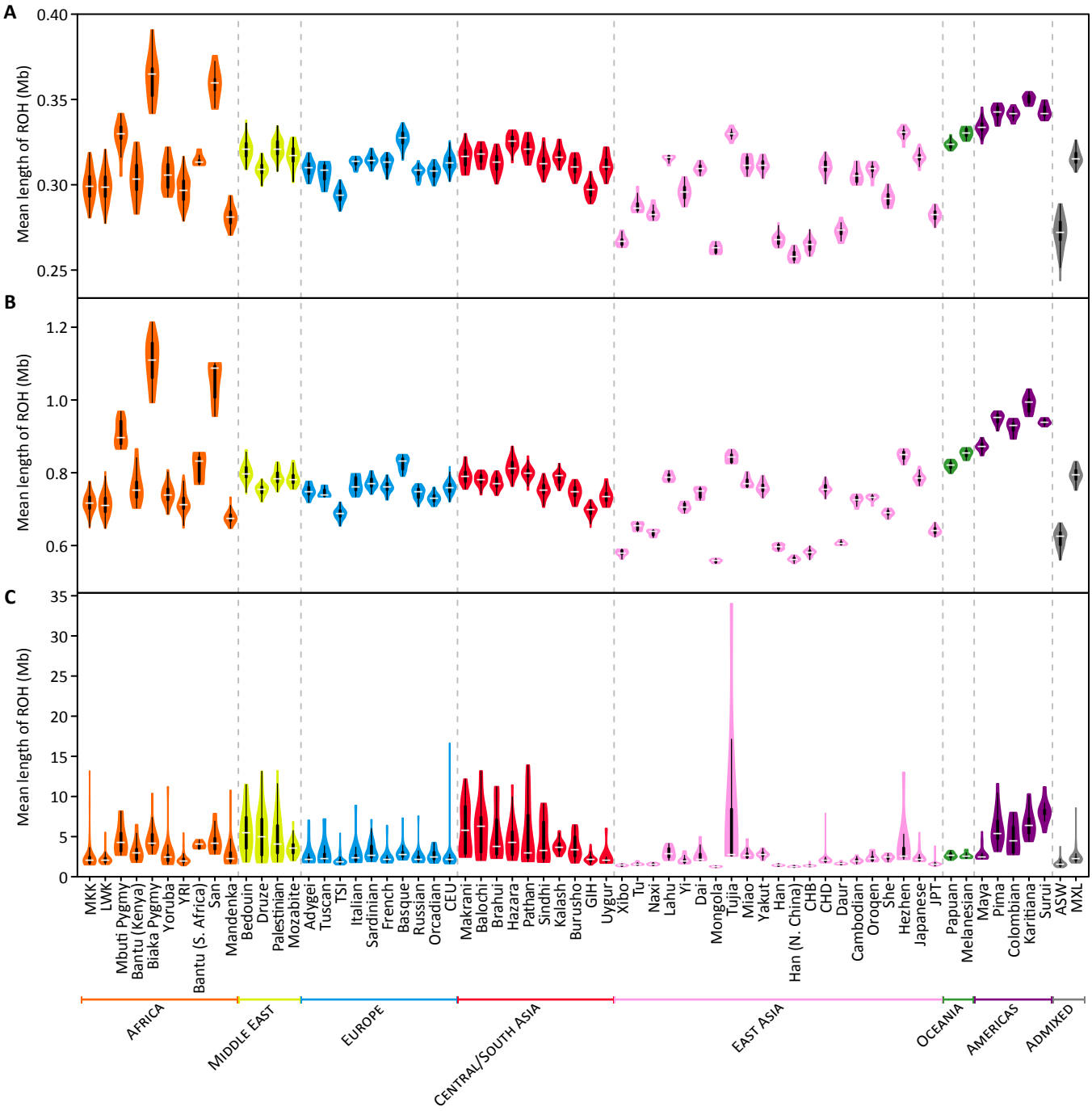

Figure S3. Mean ROH Length for Each of the Three Size Classes in Each Population

(**A**) Class A ROH, (**B**) class B ROH, and (**C**) class C ROH. Populations are ordered from left to right by geographic affiliation, and within geographic regions from left to right by increasing geographic distance from Addis Ababa, Ethiopia, and they are colored by their geographic affiliation. Vertical dashed gray lines mark the boundaries between populations from different geographic regions.

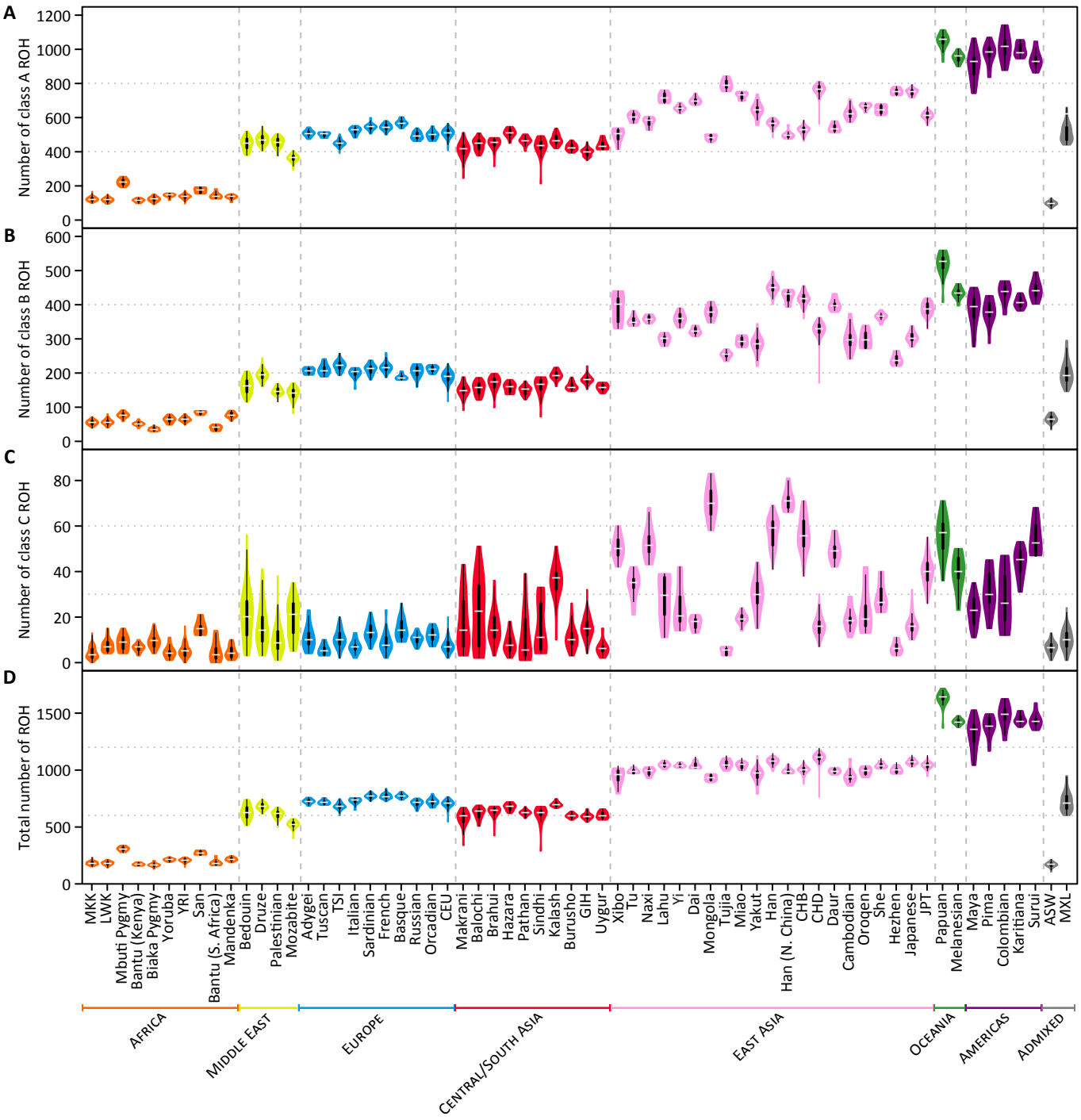

Figure S4. Total Numbers of ROH in Individual Genomes

(**A**) Total number of class A ROH, (**B**) total number of class B ROH, (**C**) total number of class C ROH, and (**D**) total number of ROH across all three classes, in each of the 53 HGDP-CEPH populations and the 11 HapMap Phase III populations. Populations are ordered from left to right by geographic affiliation, and within geographic regions from left to right by increasing geographic distance from Addis Ababa, Ethiopia. Vertical dashed gray lines mark the boundaries between populations from different geographic regions.

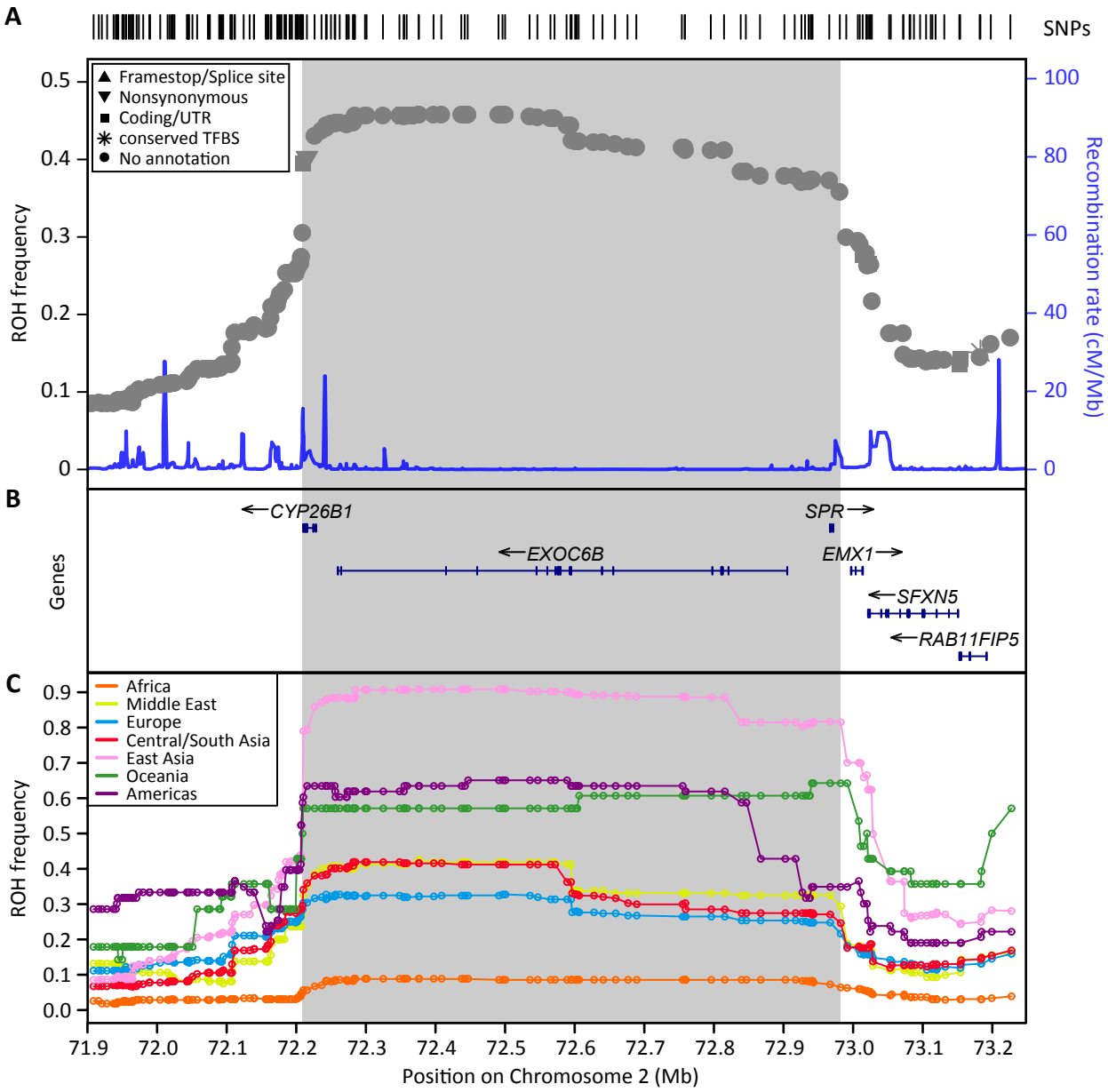

Figure S5. The Top-Ranked ROH Hot Spot on Human Autosomes

(**A**) ROH frequency calculated over all three size classes and over all populations, plotted against position on chromosome 2. The gray box indicates the location of the top-ranked ROH hotspot (**Table 1**). Population-based estimates of recombination rate are shown for release 22 of HapMap Phase II[1] (blue line). (**B**) Gene locations in release hg18 of the UCSC database.[2] (**C**) ROH frequency calculated over all three size classes, shown separately for each geographic region. Parts (**A**) and (**B**) were plotted using LocusZoom.[3] Abbreviations: UTR, untranslated region; TFBS, transcription factor binding site.

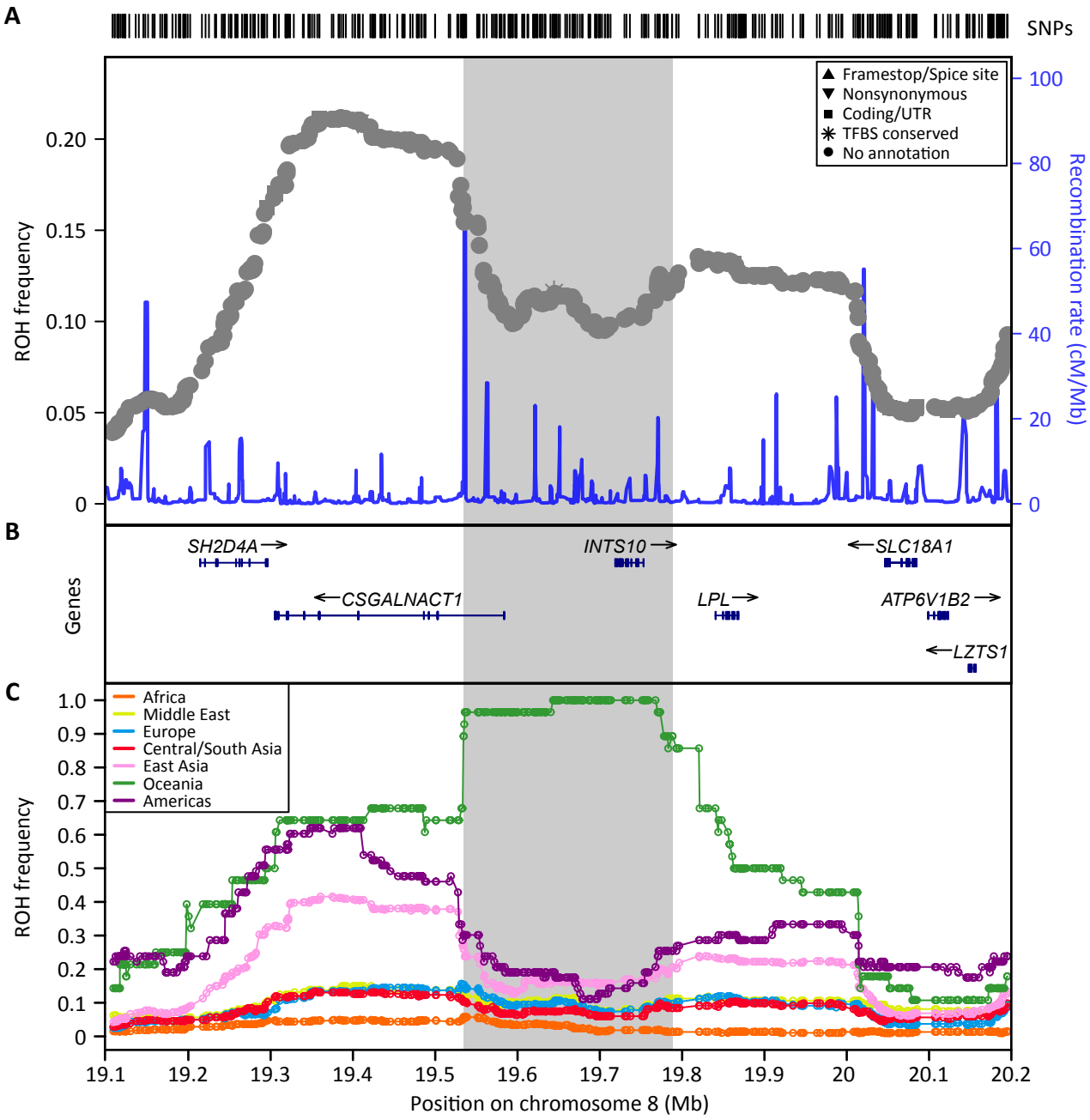

Figure S6. The Top-Ranked ROH Hot Spot on Human Autosomes in Oceanians

(**A**) ROH frequency calculated over all three size classes and over all populations, plotted against position on chromosome 8. The gray box indicates the location of the top-ranked ROH hotspot in Oceanians (**Table S6**). Population-based estimates of recombination rate are shown for release 22 of HapMap Phase II[1] (blue line). (**B**) Gene locations in release hg18 of the UCSC database.[2] (**C**) ROH frequency calculated over all three size classes, shown separately for each geographic region. Parts (**A**) and (**B**) were plotted using LocusZoom.[3] Abbreviations: UTR, untranslated region; TFBS, transcription factor binding site.

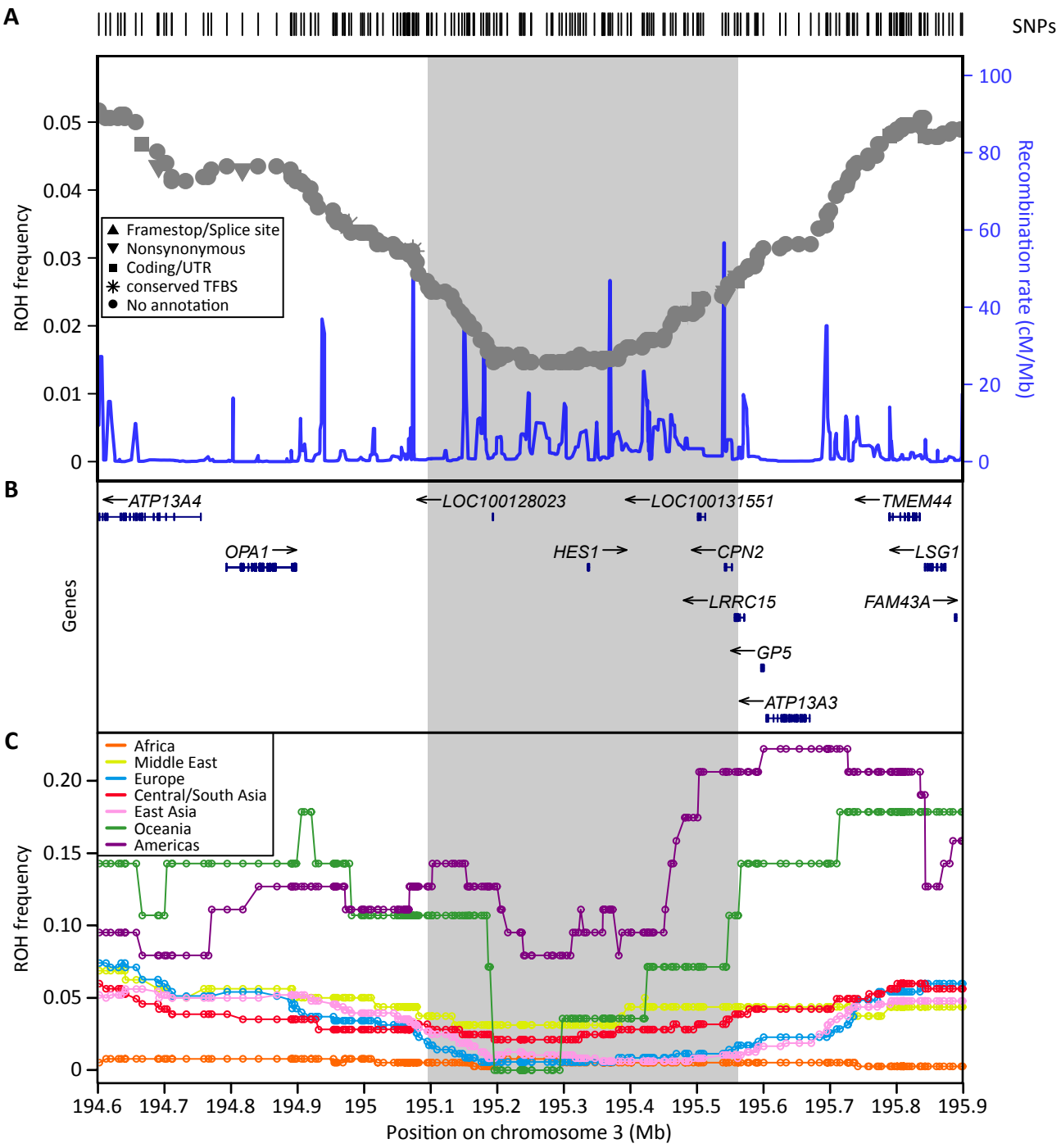

Figure S7. The Top-Ranked ROH Cold Spot on Human Autosomes

(**A**) ROH frequency calculated over all three size classes and over all populations, plotted against position on chromosome 3. The gray box indicates the location of the top-ranked ROH coldspot (**Table 2**). Population-based estimates of recombination rate are shown for release 22 of HapMap Phase II[1] (blue line). (**B**) Gene locations in release hg18 of the UCSC database.[2] (**C**) ROH frequency calculated over all three size classes, shown separately for each geographic region. Parts (**A**) and (**B**) were plotted using LocusZoom.[3] Abbreviations: UTR, untranslated region; TFBS, transcription factor binding site.

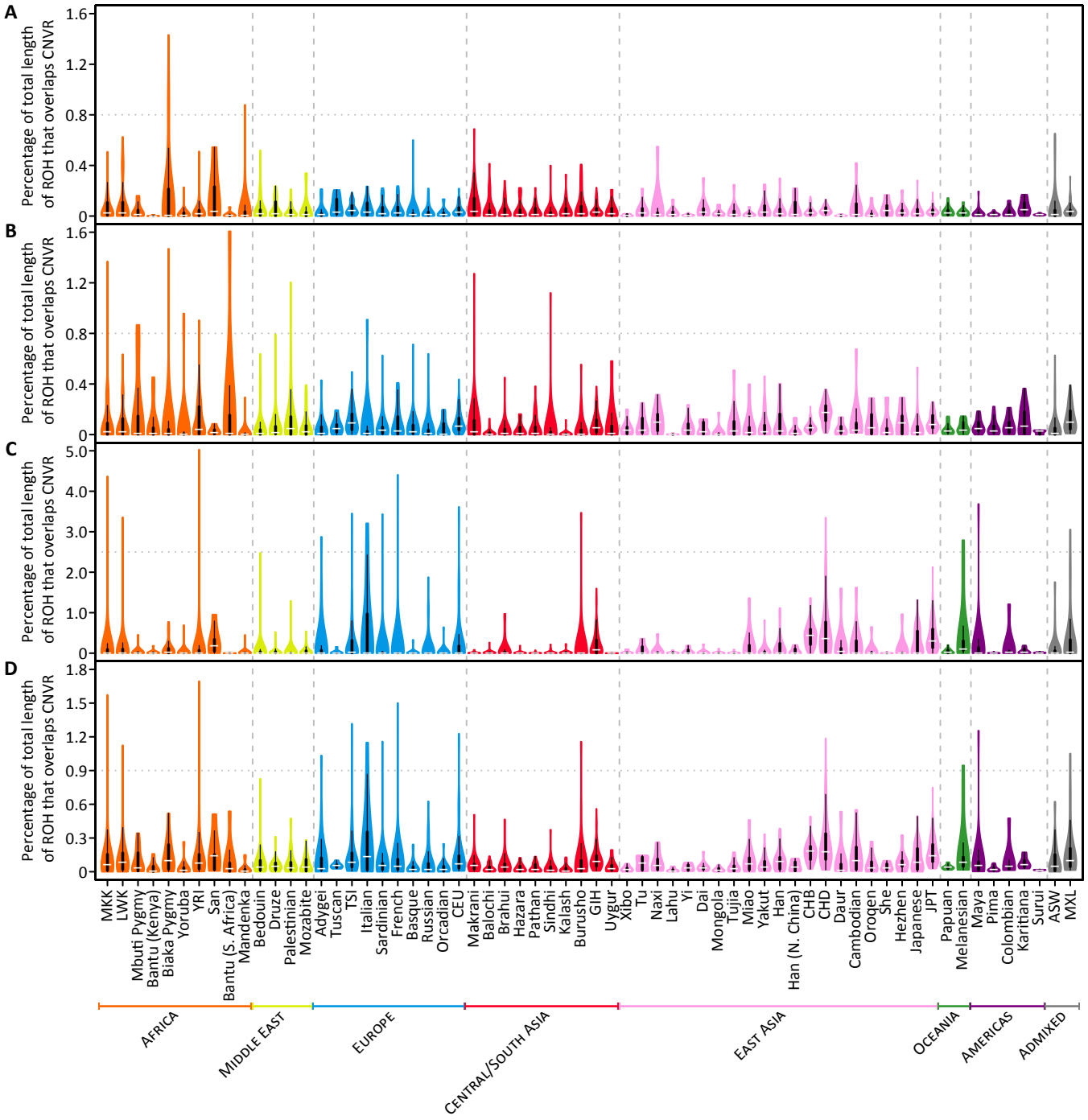

Figure S8. Overlap of ROH and CNVR in Individual Genomes

The percentage of total lengths of (**A**) class A ROH, (**B**) class B ROH, (**C**) class C ROH, and (**D**) ROH across all three classes in individual genomes that overlap CNVR identified in those genomes, in each of the 53 HGDP-CEPH populations and the 11 HapMap Phase III populations. Populations are ordered from left to right by geographic affiliation, and within geographic regions from left to right by increasing geographic distance from Addis Ababa, Ethiopia. Vertical dashed gray lines mark the boundaries between populations from different geographic regions. CNVR for the 938 individuals in the HGDP-CEPH data set were obtained from the HGDP Selection Browser (downloaded March 8th, 2010), and CNVR for the 901 individuals in the HapMap data set were obtained from Release 2 of HapMap Phase III[4] (downloaded January 27th, 2011).

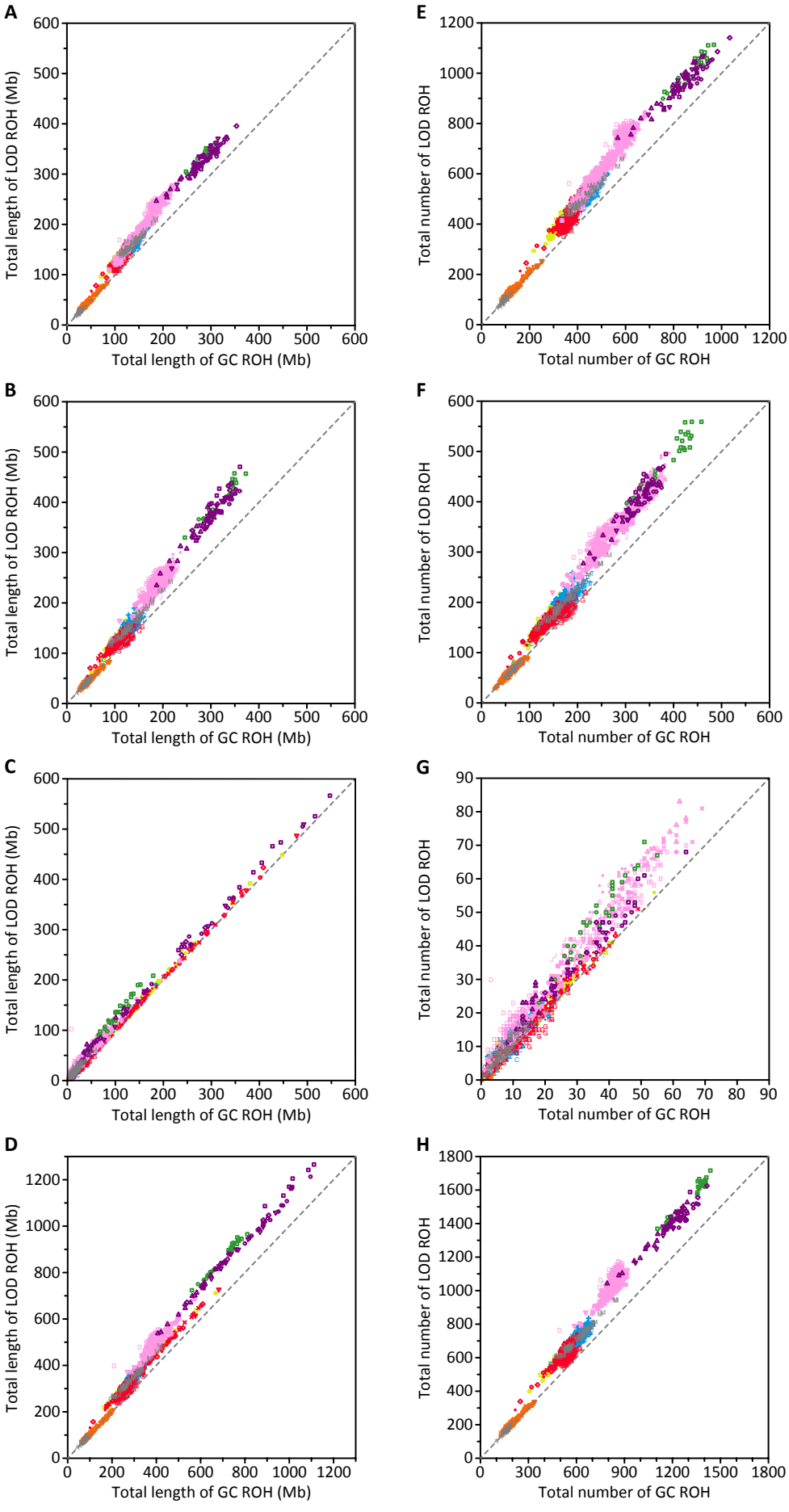

Figure S9. Comparison of per-Individual Total ROH Lengths and Numbers Detected via Likelihood-Based and Genotype-Counting Methods

Comparisons of per-individual total lengths (**A**-**D**) and total numbers (**E**-**H**) of class A ROH (**A** & **E**), class B ROH (**B** & **F**), class C ROH (**C** & **G**), and all three ROH classes combined (**D** & **H**), identified using the likelihood-based (LOD) method and the genotype counting (GC) method. Both methods were applied concurrently to the genotype data using a sliding window of 60 SNPs, and ROH were classified using the same boundary sizes (**Table S1**). When determining the homozygosity status of a window using the GC method, we used the criteria of McQuillan *et al.*:[5] ≤1 heterozygous and ≤5 missing genotypes per window, mean SNP density of ≤50kb/SNP per window, and a maximum gap between two consecutive homozygous SNPs of 100 kb. Populations are indicated by the same symbols as in **Figure 4**.

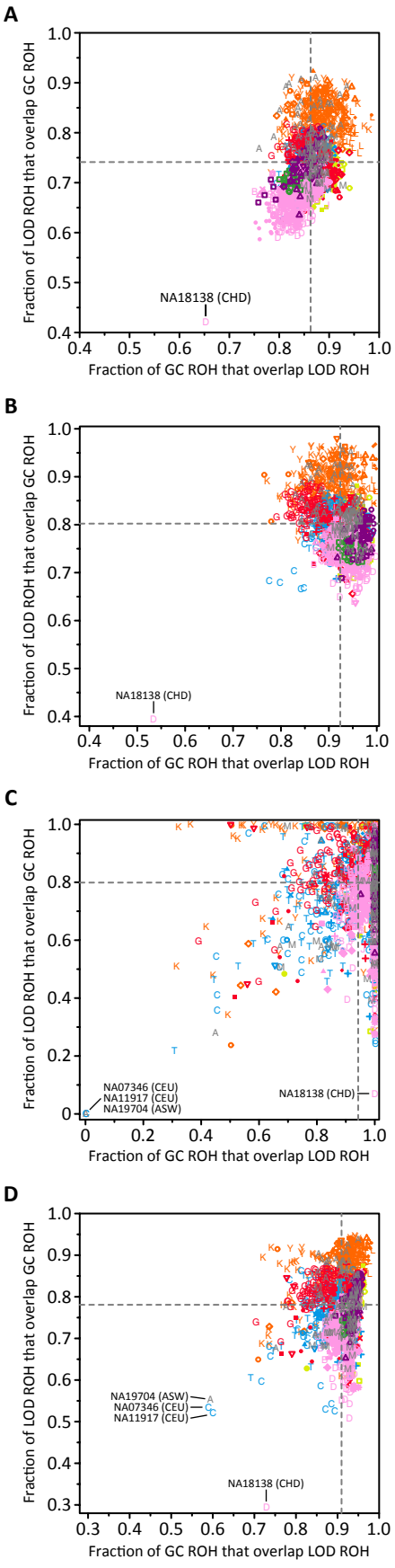

Figure S10. Overlap of ROH Detected in Individual Genomes via Likelihood-Based and Genotype-Counting Methods.

The fractions of per-individual total lengths of ROH identified using the likelihood-based (LOD) method that overlap ROH identified using a genotype-counting (GC) method, against the fractions of per-individual total lengths of ROH identified using the GC method that overlap ROH identified using the LOD method, shown separately for (**A**) class A ROH, (**B**) class B ROH, (**C**) class C ROH, and (**D**) all three ROH classes combined. Both methods were applied concurrently to the genotype data using a sliding window of 60 SNPs, and ROH were classified using the same boundary sizes (**Table S1**). When determining the homozygosity status of a window using the GC method, the criteria of McQuillan et al. were used:[5] ≤1 heterozygous and ≤5 missing genotypes per window, mean SNP density of ≤50kb/SNP per window, and a maximum gap between two consecutive homozygous SNPs of 100 kb. Populations are indicated by the same symbols as in **Figure 4**. Vertical and horizontal dashed gray lines mark the mean fraction across individuals.

Table S1. Number of Sampled Individuals, LOD-Score Thresholds for ROH Identification, and ROH-Class Boundary Sizes for Each Population

Population	Geographic region	Number of individuals	LOD score threshold	Boundary size (bp)	
				Class A & B	Class B & C
MKK	Africa	96	5.1558	486325.5	1416826
LWK	Africa	80	4.4026	481590	1315546
Mbuti Pygmy	Africa	13	2.1320	578239	1909271
Bantu (Kenya)	Africa	11	4.0622	498712	1504027
Biaka Pygmy	Africa	22	2.9786	676619	2149528
Yoruba	Africa	21	4.6202	502478.5	1528850
YRI	Africa	108	5.0270	482406	1392595.5
San	Africa	5	0.6153	526971.5	1624711.5
Bantu (S. Africa)	Africa	8	3.0647	656043.5	2095438.5
Mandenka	Africa	22	4.4230	451395	1440635
Bedouin	Middle East	45	1.5238	540741	1689397
Druze	Middle East	42	1.1807	508043	1626172
Palestinian	Middle East	46	1.6782	543776.5	1690271.5
Mozabite	Middle East	27	1.7775	528280	1547785.5
Adygei	Europe	17	0.6601	508659.5	1546061
Tuscan	Europe	7	0.4139	503582	1629045.5
TSI	Europe	88	2.5330	467503	1364957.5
Italian	Europe	12	1.1112	521248.5	1667341
Sardinian	Europe	28	1.5279	520803	1584867
French	Europe	28	0.8937	516212	1605528.5
Basque	Europe	24	1.1949	560403	1689209
Russian	Europe	25	1.8286	502367	1539373.5
Orcadian	Europe	15	1.6998	498705.5	1452138.5
CEU	Europe	107	2.4146	517859.5	1598844
Makrani	Central/South Asia	25	1.9411	530031	1687202
Balochi	Central/South Asia	24	1.4637	533668	1650771.5
Brahui	Central/South Asia	25	1.4727	521952.5	1662551.5
Hazara	Central/South Asia	22	1.2210	557995.5	1770782
Pathan	Central/South Asia	22	1.7475	545900.5	1756266.5
Sindhi	Central/South Asia	24	1.8343	517691.5	1590812.5
Kalash	Central/South Asia	23	0.5359	524842.5	1539702
Burusho	Central/South Asia	25	2.4869	514483	1570643.5
GIH	Central/South Asia	84	3.4214	475476	1336163.5
Uygur	Central/South Asia	10	2.0470	511274.5	1526191
Xibo	East Asia	9	-0.5799	394403	994858
Tu	East Asia	10	-1.1635	443416.5	1155318.5
Naxi	East Asia	8	-0.9561	433261	1078375.5
Lahu	East Asia	8	-2.2314	522309.5	1570839
Yi	East Asia	10	-1.3182	470292	1386313
Dai	East Asia	10	-0.9820	504613	1528287
Mongola	East Asia	10	-0.4183	386535.5	896699

Tujia	East Asia	10	-1.3610	571092	2050719.5
Miao	East Asia	10	-1.2605	517903.5	1556154
Yakut	East Asia	25	-0.7117	508864	1478092
Han	East Asia	34	-1.0061	399176.5	1030188.5
Han (N. China)	East Asia	10	-0.7204	379554.5	922650.5
CHB	East Asia	84	0.1686	391738.5	985513
CHD	East Asia	82	-1.5498	509342.5	1535983.5
Daur	East Asia	9	-0.9591	411640.5	1078310
Oroqen	East Asia	9	-0.5794	501657.5	1430606
Cambodian	East Asia	10	-0.7541	491889	1380519
She	East Asia	10	-1.4174	460755.5	1354487.5
Hezhen	East Asia	9	-0.8060	567854.5	1938313
Japanese	East Asia	28	-0.8764	525941	1600258.5
JPT	East Asia	86	-0.2048	434021.5	1144317
Papuan	Oceania	17	-3.2671	533037	1580777.5
Melanesian	Oceania	11	-3.5557	554474.5	1654959.5
Maya	America	21	-2.1264	562835.5	1799162.5
Pima	America	14	-3.4241	597895	2137062
Colombian	America	7	-3.6430	585419.5	2067894.5
Karitiana	America	13	-4.0559	615621.5	2191780.5
Surui	America	8	-6.5691	586472.5	2115337
ASW	Admixed	40	5.4990	414633.5	1129271
MXL	Admixed	46	1.7014	526641	1623921

The ranges of the boundaries are the shaded areas of **Figure 2C**.

Table S6. The Top-Ranked ROH Hot Spot on Human Autosomes in Each Geographic Region

Geographic region	Chr	Genomic region (kb)			ROH frequency			Content
		Begin	End	Length	Max	Mean	St Dev	RefSeq Genes
Africa	18	64,153	64,250	97	0.187	0.179	0.011	-
Middle East[a]	15	69,852	70,666	814	0.563	0.510	0.068	*ARIH1*, *C15orf34*, *CELF6*, *GRAMD2*, *HEXA*,[b] *MYO9A*, *NR2E3*, *PARP6*, *PKM2*, *SENP8*, *THSD4*, *TMEM202*
Europe[a]	15	69,858	70,640	782	0.521	0.485	0.054	
Central/South Asia[a]	15	69,863	70,666	803	0.479	0.449	0.034	
East Asia[c, d]	2	72,210	73,026	817	0.909	0.854	0.074	*CYP26B1*, *EMX1*, *EXOC6B*, *SFXN5*, *SPR*
Oceania[d]	8	19,534	19,788	254	1	0.972	0.031	*CSGALNACT1*, *INTS10*
Americas[d]	2	177,856	178,672	815	0.984	0.939	0.031	*AGPS*,[e] ***PDE11A***,[f] *TTC30A*, *TTC30B*, *LOC100130691*

Genes in **bold** and underlined are associated with autosomal dominant and autosomal recessive diseases, respectively, in the OMIM database.

[a]These regions overlap the eighth-ranked ROH hotspots in the full data set (**Table 1**).

[b]Tay-Sachs disease (MIM: 272800).

[c]This region overlaps the top-ranked ROH hotspot in the full data set (**Table 1**).

[d]These hotspots overlap regions identified in previous genomic surveys as probable targets of recent positive selection.[6-9]

[e]Rhizomelic chondrodysplasia punctate (MIM: 600121).

[f]Cushing syndrome (MIM: 610475).

Table S7. The Top-Ranked ROH Cold Spot on Human Autosomes in Each Geographic Region

Geographic region	Chr	Genomic region (kb)			ROH frequency			Content
		Begin	End	Length	Min	Mean	St Dev	RefSeq Genes
Africa	15	88,175	88,541	366	0	0	0.001	*AP3S2*, *IDH2*, *SEMA4B*, *ZNF710*
Middle East	11	3,165	3,596	431	0.006	0.013	0.005	*C11orf36*, *LOC650368*, *MRGPRE*, *MRGPRG*, *ZNF195*
Europe	18	3,436	3,606	170	0.003	0.004	0.002	*DLGAP1*, *FLJ35776*, ***TGIF1***[a]
Central/South Asia	6	107,300	107,515	216	0.004	0.011	0.006	*BEND3*, *C6orf203*, *LOC100422737*
East Asia[b]	3	195,143	195,672	529	0.006	0.010	0.004	*ATP13A3*, *CPN2*, *GP5*, *HES1*, *LRRC15*, *LOC647323*, *LOC100128023*, *LOC100131551*, *LRRC15*
Oceania	3	177,155	177,835	681	0	0.004	0.011	-
Americas	3	126,174	126,318	144	0.016	0.043	0.028	*HEG1*, *SLC12A8*

Genes in **bold** are associated with autosomal dominant diseases in the OMIM database.

[a]Holoprosencephaly (MIM:142946).

[b]This region overlaps the top-ranked ROH coldspot in the full data set (**Table 2**).

Additional Supplemental Tables

Table S2. The Frequency of Class A ROH at Each SNP in the Combined Data Set

ROH frequencies are given separately for each geographic region, as well as across all individuals in the data set.

Table S3. The Frequency of Class B ROH at Each SNP in the Combined Data Set

ROH frequencies are given separately for each geographic region, as well as across all individuals in the data set.

Table S4. The Frequency of Class C ROH at Each SNP in the Combined Data Set

ROH frequencies are given separately for each geographic region, as well as across all individuals in the data set.

Table S5. The Frequency of ROH Calculated over All Three Size Classes at Each SNP in the Combined Data Set

ROH frequencies are given separately for each geographic region, as well as across all individuals in the data set.

Supplemental References

1. The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.
2. Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., et al. (2010). The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38, D613-619.
3. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336-2337.
4. The International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467, 52-58.
5. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., et al. (2008). Runs of homozygosity in European populations. Am J Hum Genet 83, 359-372.
6. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913-918.
7. Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet 3, e90.
8. Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W., and Pritchard, J.K. (2009). The role of geography in human adaptation. PLoS Genet 5, e1000500.
9. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19, 826-837.