A second generation human haplotype map of over 3.1 million SNPs

The International HapMap Consortium[1]

Supplementary material

[1] See end of manuscript for Consortium details

Supplementary text 1. The density of common SNPs in the Phase II HapMap and the assembled human genome.

To estimate the fraction of all common variants on the autosomes that have been successfully genotyped in the consensus Phase II HapMap we note that in YRI (release 21) there are 2,334,980 SNPs with MAF≥0.05. Across the autosomes, the completed reference sequence assembled in contigs is 2.68 billion bp. Assuming that the allele frequency distribution in the YRI is well approximated by that of a simple coalescent model and using an estimate of the population mutation rate of $\theta = 1.2$ per kb for African populations[1,2] the expected number of variants with MAF≥0.05 in a sample of 120 chromosomes is

$$E(S_{MAF \geq 5\%}) = L\theta \sum_{i=6}^{114} 1/i$$

where L is the total length of the sequence[3]. Using the values above we expect 9.7 million common SNPs in the sample. We therefore estimate that 24% of all common variants are present in the Phase II HapMap. For the other analysis panels a model of constant population size is not appropriate, but is nevertheless instructive. Using an estimate of the population mutation rate of $\theta = 0.8$ per kb for both non-African panels[1,2] we estimate that 32% of all common SNPs in CEU and 29% of all common SNPs in CHB+JPT are present in the Phase II HapMap. Because diversity in non-African populations is typically biased away from low-frequency variants[1,2], the estimates in non-African populations are probably underestimates of the proportion of common SNPs in HapMap Phase II.

Previously, we estimated that approximately 70% of all SNPs with MAF≥0.05 in YRI were present in dbSNP release 125[4]. Given that assays could be designed for approximately 61% of all SNPs in dbSNP release 122, 62% of all submissions passed QC and 91% of submissions that were QC+ in one panel but not three, we would therefore expect approximately 70 x 0.61 x 0.62 x 0.91 = 24% of all common SNPs to be QC+ in YRI. The agreement between estimates is remarkable.

Supplementary text 2. Analysis of data quality

2.1 Analysis of the relation of amplicon structure to genotyping error

An important aspect of experimental design for the additional SNPs genotyped for the Phase II HapMap is the amplicon long-range PCR structure of the Perlegen design. Undetected polymorphism in the primer regions, non-uniqueness in primer mapping or errors in the genome assembly can create different types of problem for such a design. Importantly, such problems will lead to clustering of errors within the genome, which might have potentially important effects for downstream analyses.

Details of the amplicons primers used in the construction of the Phase II HapMap and their mapping to NCBI Build 35.1 are available from http://genome.perlegen.com/pcr/ in the file `PP_BLAST_B35.dat` and also from the HapMap website. On the HapMap web-site, mapping of rs ID and assay ID to amplicon and the Phase II HapMap data sets is available from the file `perlegen_amplicon_assaylsid_mapping_rel21.txt`, while summary information for amplicon quality and BLAST hits for each amplicon is available in the file `perlegen_amplicon_summary_rel21.txt`. One caveat identified in joining the Perlegen amplicon mapping data with the HapMap data set tables is that some rsids differ between the two sets of tables.. In this case, assaylsid proved more reliable in performing this join. Of 302,920 amplicons, 296,273 uniquely mapped to NCBI Build 35.1, 74 had no Build 35.1 coordinates, 2,774 had multiple inter-chromosomal hits, and 3,799 had multiple intra-chromosomal hits. No filtering was performed based on non-unique mapping to the human reference genome. For the uniquely mapping amplicons, the mean length was 8.8 kb with a range of 619 bp to 23.8 kb. The mean number of SNPs in an amplicon was 14.6 with a range from 1 to 743 (excluding the 976 with no SNPs in Phase II).

To summarise amplicon quality, we derived a simple metric, the amplicon quality score (AQS), which is the proportion of Perlegen assayed SNPs in the amplicon that passed HapMap quality control measures (QC+) in all three analysis panels, extracting data from the redundant un-filtered data set. Supplementary Figure 1 shows a frequency histogram of AQS for the amplicons with release 21 assays.

We identified 472,710 rsids with duplicate non-Perlegen (NPRL) and Perlegen (PRL) assays, of which 316,362 were QC+ across all three panels in both NPRL and PRL. From the PRL set of SNPs, we tallied each matched genotype for genotype:genotype accordance, genotype:genotype discordance, or genotype:no call discordance. Of the resulting 85,417,740 genotype comparisons, we observed 83,008,843 accordant genotypes (accordance rate 97.18%), with 433,838 discordant genotypes (discordance rate 0.51%). There were 1,449,972 where a genotype was called by NPRL with a no call by PRL (NPRL/PRL genotype:no nocall discordance rate 1.70%), while there were 525,087 in the opposite direction (PRL/NPRL genotype:no call discordance rate 0.61%).

Of the SNPs described above, 303,660 were assignable to a single amplicon. Supplementary Figure 2 shows discordance plotted against the reference allele frequency from the NPRL assay and a summary of the results is shown in Supplementary Table 1. For our analysis, “reference allele” refers to the allele of lower

alphabetic order. Genotype discrepancies are strongly enriched among low quality amplicons and are largely driven by SNPs identified as monomorphic by the Perlegen assay and polymorphic by the other assay. However, over 90% of all SNPs lie in amplicons with AQS≥0.5 for which the discordance rate is less than 0.5%. Furthermore, discordance in high quality amplicons does not appear to be driven by apparently monomorphic SNPs in the Perlegen assay. Together these results indicate that knowledge of the amplicon structure can provide a powerful source of information to help identify genotypes of poor quality. For example, exclusion of SNPs with Perlegen assay reported MAF<5% in amplicons with AQS < 0.4 would remove the majority of discrepancies.

2.2. Analysis of genotype discordance from overlap with SeattleSNPs

Seattle SNPs genotypes were obtained by targeted sequencing of genic regions in 22 or 23 individuals of European decent. HapMap genotype data came from probe based genotyping of the 60 CEPH founders. A subset of the subjects was genotyped in both groups on 1,828 SNPs. The number of subjects varied, but ranged between 5 and 23 individuals. We identified 103 SNPs 5.6 % for which at least one individual's genotype was called differently by HapMap and SeattleSNPs. Of these, 68 SNPs have a single discrepant subject, 19 have 2 discrepant subjects, 1 has 3, and 15 have 4 or more discrepancies. Of the 38,453 opportunities to detect discrepancies, we found 258 discrepant genotypes (0.7%). We summarized the genotype differences in Supplementary Table 2. In addition, allele frequencies at 2,932 SNPs identified as polymorphic in either SeattleSNPs or HapMap Phase II were compared. Overall, we find little evidence for significant differences in allele frequency; compared to the 29, 2.9 and 0.29 SNPs we expect to be significant at the 0.01, 0.001, and 0.0001 level, we observe 12, 2, and 9 respectively. Therefore we only observe an excess of SNPs showing very strong allele frequency differences and in all nine cases HapMap Phase II reports the SNP to be almost or completely monomorphic.

2.3 Analysis of genotype discordance from fosmid end sequences

Nine HapMap individuals were selected for fosmid end sequencing for the Human Genome Structural Variation project. The 7 sets that are complete or nearly complete (0.4X sequence coverage per individual, with 0.8X for NA18507) were selected for further analysis. Using ssahaSNP[5], SNPs were detected from reads relative to the reference sequence. None of this discovery was submitted to dbSNP prior to any part of HapMap Phase II, thus making it an independent data source. If variants (i.e. non-reference alleles) were identified at a SNP successfully typed on the same individual in HapMap Phase II the genotype is marked as either concordant if it also carries at least one non-reference allele or discordant if it is reported as homozygous for the reference allele. Counts in each class are shown in Supplementary Table 3. Most platforms show similar levels of discordance, from 0.5 – 2%. Note that because discrepancies can only be detected in individuals carrying the non-reference allele, which is also likely to have a higher error rate through undetected polymorphism in LD in nearby primer regions, the average error rate is likely to be lower.

Supplementary Table 3 shows that the Infinium assay from Illumina has a very low discrepancy rate. To gain a better understanding of the cause of discrepancies genotype calls were compared against the Infinium assay on two individuals (NA18507 and NA18555) and the nature of any discrepancy was noted. Across all platforms (and particularly for the Perlegen platform) the single greatest form of discrepancy was when the Infinium assay reported a heterozygote and the alternative platform reported a homozygote for the reference allele (32% of all discrepancies overall, 45% of all discrepancies with Perlegen genotypes). Less than 10% of all discrepancies were caused by reports of homozygous reference allele by one platform and homozygous alternative allele by the other.

2.4 Analysis of monomorphism/polymorphism discrepancies

The above analyses suggest that a significant contribution to the genotype error structure comes from SNPs falsely identified as monomorphic on one platform. To further address this issue we compared all QC+ SNP submissions across platforms and centres to identify duplicate SNP submissions that were identified as polymorphic on one platform and monomorphic on another (also excluding submissions with more than five missing data calls). Results comparing each platform to Infinium are shown in Supplementary Table 4. Overall, we find that 0.09% of SNPs show discrepancies in mono/polymorphism status and that platforms differ in the rate of such occurrences. In the majority of cases discrepancies relate to SNPs for which the minor allele frequency is less than 10%. In addition, for most platforms we observe an excess of monomorphism calls compared to the Infinium assay. Another important finding is that we identify a small, but potentially important, fraction (0.02%) of SNPs where platforms agree on monomorphism, but of different alleles. These are not particularly biased towards cases that may be due to errors in reporting the strand (i.e. A/T and G/C SNPs) and may reflect problems in SNP localization, errors in informatics, or difficulties in assay design and calling (data not shown).

2.5 Interchomosomal LD

Incorrect mapping of SNPs to genomic location can potentially lead to inconsistencies in local patterns of LD. To assess the evidence for mis-mapping we searched each analysis panel for SNPs with MAF of at least 0.05 that have an r^2 of at least 0.8 to another such SNP on a different chromosome. Although it is possible for such inter-chromosomal LD to arise from strong epistatic selection, the most likely explanation is incorrect SNP mapping. In each analysis panel 2,000-3,000 such SNP pairs (approximately 0.1% of all SNPs) were identified. In the majority of cases one of the SNPs also showed no strong LD to other SNPs within the same mapped region, further suggesting that these are the result of mis-mapping. Among the minority of SNPs that show both inter-chromosomal LD and local LD 87% occur in segmental duplications (compared to 2% overall). Particularly notable are the clusters of SNPs with r^2 =1 on chromosomes 1, 13 and 15 that overlap an annotated duplication on the Y chromosome (see Supplementary Figure 3). However, only a small fraction of the 2% of all SNPs mapping to annotated segmental duplications show evidence of inter-chromosomal LD. A list of SNPs showing inter-chromosomal LD is available for bulk download from the HapMap web site.

Supplementary Text 3. Analysis of population stratification

For these analyses, we filtered out SNPs with less than 99% complete genotyping, and removed a single JPT individual (NA19012) with less than 90% genotyping, leaving over 2 million SNPs. For each pair of individuals (269 individuals, 36046 pairs) we calculated the proportion of SNP alleles shared identical-by-state (IBS); a summary of this raw pairwise measure of genetic similarity is shown in Supplementary Table 6. The IBS metric ranges from 0.77 (a between population comparison) to 0.90 (a CEU parent-offspring pair). Looking within subpopulations, and ignoring parent-offspring pairs, all CHB and JPT individuals are more similar to each other than any two CEU individuals are to each other, who are, in turn, more similar to each other than any two YRI individuals are to each other. Considering individuals from different populations: CHB and JPT are more similar to each other than CEU and YRI are to themselves (the distribution for CHB/JPT pairs virtually overlaps the CHB/CHB and JPT/JPT distributions). CEU/CHB and CEU/JPT pairs are only slightly less distant than YRI/YRI pairs. YRI to non-YRI comparisons consistently show the lowest levels of IBS.

We also applied principal components analysis methods[6] to detect population stratification. In some of these analyses, a small number of outlier samples, which could represent genetic outliers or (more likely) cryptically related samples, were detected and removed. From the analysis of all 209 founder samples the top two principal components are highly statistically significant (P-value < 1e-12) and clearly separate the three analysis panels, as expected. Analyzing each analysis panel separately, no evidence for further substructure was detected in either CEU or YRI, with the top principal components not being statistically significant. In an analysis of 89 CHB + JPT samples the top principal component is highly statistically significant (P-value < 1e-12) and clearly separates CHB from JPT. One JPT sample, NA18976, appears to have mixed ancestry. The second principal component is significant (P-value = 0.006) and is more varied for CHB than JPT, suggesting population structure in CHB. Indeed, analysis of 44 JPT samples shows no significant population structure but analysis of 45 CHB samples shows a significant top principal component (P-value = 0.002) which is strongly correlated (Pearson correlation coefficient = 0.93) in CHB samples to the second principal component of 89 CHB + JPT samples. The level of population structure in CHB is equivalent to what one would see with two discrete subpopulations with FST = 0.002. This is a smaller effect than the FST = 0.007 between CHB and JPT. Note, however, that FST can vary substantially along a genome[7].

Supplementary Text 4. Analysis of relatedness

Within each population, we next estimated the genome-wide level of relatedness between all pairs of individuals. We use a simple method of moments approach[8] to estimate the probability of sharing Z=0, 1 or 2 alleles identical-by-descent (IBD) for any two individuals from the same homogeneous, random-mating populations, and also π, the proportion of alleles shared IBD between two individuals, as $P(Z = 1)/2 + P(Z = 2)$.

As previously reported[4], we observed close relationships between individuals in the YRI and CEU populations. In particular, NA18913 and NA19238 (YRI) are a parent-offspring pair (with estimated IBD probabilities of 0.01, 0.98 and 0.01 for sharing 0, 1 and 2 alleles IBD); also, NA19130 and NA19192 (YRI); NA19092 and NA19101 (YRI) are cousins. The elevated level of relatedness between the other known blood relatives of these individuals was consistent with these relationships inferred from the genetic data. A number of CEU individuals show higher than expected relatedness also.

The estimates assume a homogeneous, random-mating population and are not constrained to biologically plausible values, to yield more unbiased results (i.e. sharing could be estimated as negative). Although the precise values are likely less accurate for very distantly related pairs, the general conclusion that a significant proportion of pairs show low but non-zero levels of relatedness is also supported by the segmental sharing analyses.

Supplementary Text 5. Segmental analysis of relatedness

We searched for extended segments shared between individuals in the same analysis panel. Specifically, we used a hidden Markov Model (HMM) approach to provide multipoint estimates for each pair of individuals sharing either 0, 1 or 2 alleles identical-by-descent (IBD) at a particular position given the observed pattern of IBS sharing[8]. Within each analysis panel, all pairs with at least some degree of estimated genome-wide relatedness were included in analysis; pairs showing close relationships (in CEU and YRI) were excluded from these analyses, as were a small number of individuals based on the stratification analyses. As it stands, the HMM requires that SNPs are in approximately linkage equilibrium at the sample level: we therefore pruned the list of SNPs to remove local LD within each analysis panel. We then formed a consensus set of SNPs that, within each analysis panel, were polymorphic, showed low levels of missing data and were in approximate linkage equilibrium. The final SNP set consisted of 45,240 autosomal SNPs (an average inter-SNP distance of 60kb). This restricted, consensus set was selected so that rates of background LD and SNP density were similar between analysis panels.

Although this SNP density is easily sufficient to detect longer segments, smaller segments will be harder to detect and the boundaries of segments will be less well resolved. For the three pairs in Figure 3, comparing the total length of segments called versus the genome-wide estimates of relatedness suggests that segments were under-called for the most distantly related pair. In other words, and as one might expect, smaller segments between more distantly-related individuals are harder to detect. Nonetheless, the principle we prove here is that this kind of SNP data can reveal extended, recent sharing in general populations, over and above background LD. In as much as the focus is on more recent, rarer variation, it should be noted that such segments will also tend to be longer and therefore easier to detect.

We also investigated the relationship between “rare variation” and segmental sharing as follows. We identified all SNPs with complete genotyping that showed only two copies of the rare allele in two heterozygous founders in each population. These instances of SNP/pair combinations we call “two-SNPs”. We can then ask what proportion of two-SNPs fall within a shared segment of IBD. Population genetic theory states that rarer SNPs are more likely to be recent and therefore it is more likely that two copies of the same recent, rare variant sit on similar local chromosomal backgrounds. Table 5 shows the number of two-SNPs identified in each population and the proportion that fall in shared segments versus what we would expect by chance. If we take the total length of the genome spanned by autosomal HapMap Phase II SNPs to be 2,782Mb, we can use the proportion of pairs of genomes covered by shared segments to give the expected proportion of two-SNPs that would be fall in shared segments if there were no relation between rare variation and extended segmental sharing. We see approximately a 7-fold increase in the number of two-SNPs within shared segments compared to chance, which strongly suggests that extended shared segments do indeed track shared rare variation. It is important to note that a two-SNP is only a weak proxy for rarer variation (i.e. 2 out

of 120 alleles is not in fact particularly rare, and the population frequency will often be substantially higher) and so this analysis undoubtedly underestimates the true relation between rare variation and extended sharing.

Supplementary Text 6. Analysis of homozygosity

Identifying contiguous runs of homozygous SNPs

For each run of consecutive homozygous genotype calls, the homozygous probability score (HPS) was calculated from the product of the observed homozygosity within an analysis panel for each SNP in a detected homozygous segment. Segments were not allowed to cross centromere or contig boundaries as well as inter-SNP distances greater than 13kb; this latter cutoff allows inclusion of approximately 99.9% of all neighbouring SNPs that do not reside on contig boundaries. Allowing segments to span contigs as well as using much longer inter-SNP cutoffs might produce spurious calls of homozygous segments in regions of low SNP density. To additionally account for regions of low SNP density, segments were also filtered to have a SNP density of greater than 0.2 SNP/kb. Significant stretches of homozygosity were identified as those with an HPS score ≤ 0.01.

After removal of putative deletions as described below, we found extensive stretches of homozygosity in all individuals and on all chromosomes. Based on the above parameters, average genome-wide coverage by homozygous segments in YRI: 660 Mb (22.0x103 segments; 8.0x105 SNPs), CEU: 950 Mb (18.9x103 segments; 11.2x105 SNPs), CHB: 1,020 Mb (17.3x103 segments; 12.1x105 SNPs), JPT: 1,030 Mb (17.2x103 segments; 12.2x105 SNPs). To more extensively filter out segments that might be attributed to simple identity by state, we calculated a length cutoff that would be inclusive across all samples and chromosomes by determining the maximum length segment for each individual and chromosome and then picking the lowest maximum length segment that was observed. This value of 106 kb for the current dataset was used to filter data as summarized in Table 5 and Figure 3.

One caveat that should be considered in understanding this dataset is that due to the high SNP density, even a low homozygote-to-heterozygote error rate of 0.2% means that on average, every 500 SNPs there could be an errant heterozygote genotyped in an otherwise contiguous region of homozygosity. To more fully account for putative autozygous segments, we search for sampled chromosomes that exhibited excess homozygosity with respect to the distribution observed for a particular analysis panel. In brief, we first identified the lowest maximum length segment for each analysis panel and chromosome to allow inclusion of all samples from each panel, while appreciably trimming small segments that were more likely to represent localized LD. The total length of homozygous segments larger than this cutoff was calculated for each sampled chromosome, following which we used a dynamic programming algorithm to remove any extreme outlier samples and calculated the mean and standard deviation for each chromosome for each analysis panel from the remaining samples. Chromosomes with excessive homozygosity were defined as those that were greater than 2 SD from the mean of that chromosome for their respective analysis panel. A total of 225 chromosomes were selected for further analysis (YRI=83, JPT+CHB=73, CEU=69).

Following this, we concatenated adjoining segments and segments separated by one or two heterozygotes. This data was subsequently filtered for regions that possessed a SNP density of at least 1 SNP every 5 kb and length greater than 3 Mb. Supplementary Table 7 shows for each subject group those samples

that had multiple non-adjoining regions of putative autozygosity. Of special note are JPT subjects, NA18987 and NA18992, each of which had nine such regions on seven different chromosomes; the total length of these regions on NA18987 was approximately 118 Mb while NA18992 had approximately 165 Mb. Supplementary Table 8 shows data for subjects that had only one region greater than 3 Mb. Both tables provide the endpoints of the concatenated regions, the region's length, the number of genotyped SNPs, and the number of those SNPs that were heterozygous.

Removal of putative hemizygous deletions

One potential confounder in detecting homozygous segments is hemizygous deletions, which may also appear as contiguous runs of homozygous genotypes. Because of this possibility, we developed a systematic process to find the intersections of homozygous segments with potential deleted regions at both the global and sample levels.

At a global level, we found the intersection with regions that commonly experience somatic deletions in lymphoblastoid cell lines: IgH, IgLV, or IgKV immunoglobulin gene clusters (chr2, 88.9-90.0 MB; chr14, 105.2-106.4 MB; chr22, 20.7-21.6 MB), as well as with copy number variable (CNV) regions identified on the 500K EA platform[9] with combined gains and losses > 10 (n=90; Supplementary Table 11C in reference 9). Chromosomal abnormalities can potentially skew genotypes across long portions of chromosomes and may represent LOH. Previously detected chromosomal abnormalities in the HapMap samples[9] (Supplementary Tables 5C, 5D, 5E in reference 9) were examined for strong or weak chromosomal loss (i.e. deleted in all/most cells versus only in a small percentage of cells) and assessed for the proportion of heterozygote and null genotypes. Abnormalities were considered putative deletions if they possessed <15% heterozygote calls and >5% null genotypes. In addition to these abnormalities, we imported the sample level CNV calls from the Affymetrix 500K EA platform[9] (Supplementary Table 10 in reference 9), and intersected homozygous segments with regions identified as a sample level "loss" (n=3,442)

To more extensively account for deletions, we downloaded the raw Affymetrix 500k data from the HapMap web-site. dChip was used to perform normalization, combining of sub-arrays, and modelling using standard settings for copy number analysis (http://www.dchip.org). Copy numbers were inferred with median smoothing and a window of 10 SNPs, and the values were exported into our database. We ordered each individual's genome-wide data, trimmed 10% from the high and low ends, and determined the mean and standard deviation of the remaining values for each chromosome. Regions of SNPs with contiguous decreased copy number values greater than four standard deviations from the mean were marked for further investigation, and neighboring regions with>=50 loci were concatenated if separated by <=10 loci. Regions were filtered for those with>4 loci, the proportion of heterozygote genotypes for that individual in the HapMap Phase II consensus dataset determined, and those with less than 15% heterozygous genotypes considered as putative deletions. 33,754 regions were detected. A typical individual had an average of 125 regions that covered between 3 Mb to 7 Mb of the genome.

If a homozygous segment intersected with multiple deletions, the highest and lowest boundaries across them were used. If a homozygous segment intersected incompletely with these combined regions, the remaining non-intersecting sub-segments were placed back into the analysis.

Supplementary Text 7. Perlegen genotyping protocols

Amplicon primer design

Long-range PCR assays were designed using OLIGO primer design software (Molecular Biology Insights). Primers were selected to have similar stringency and to map uniquely to NCBI Build 33. From a collection of all suitable candidate primers with amplicon lengths between 3 kb and 12 kb, custom software was used to select a minimum spanning set having maximum coverage with minimal overlap between adjacent amplicons. For the development of the Perlegen haplotype map[10], 293,061 primer pairs had been designed using these criteria; these plus 13,075 new primer pairs chosen to cover SNPs not covered by that set were used. The amplicons resulting from the 306,136 primer pairs had a median length of 9.5 kb. These primers were multiplexed to 11 or 12 primer pairs per reaction, distributed to avoid unwanted amplification products. The primer pairs as designed together amplified a total of 2.6 billion base pairs of genomic sequence.

DNA amplification

Multiplex long range PCR reactions were set up as follows (per reaction): 11 ng of genomic DNA was amplified using 11-12 PCR primer pairs (0.16µM of each primer), 0.29 U EpiTaq (Epicentre), 0.1 µg TaqStart antibody (Becton Dickinson), 0.31 µl Antibody buffer, 2.25 mM dNTPs , 0.14 µl Tricine (1 M), 0.17 µl DMSO, 22 mM Tris-HCl (pH 9.1), 1.2 mM $MgCl_2$, 6 mM ammonium sulfate, 2.6 mM KCl, and 0.25 µl 10× MasterAmp PCR enhancer (Epicentre), in a volume of 6 µl. Thermocycling was performed using a 9700 cycler (Perkin-Elmer) as follows: initial denaturation for 3 minutes at 94°C, 10 cycles of (94°C 2 s, 64°C 15 minutes per cycle), 28 cycles of (94°C 2 s, 64°C 15 minutes with a 20 s increase per cycle), then a final 60 minute extension at 62°C.

DNA labeling and hybridization

For each of the 49 high-density oligonucleotide arrays, corresponding PCR products were combined into one tube per individual and purified using the Montage PCR clean up kit (Millipore). The pooled purified PCR products were then adjusted to 1.8 µg/µl and 50 ug was incubated for 8 minutes at 37°C with 0.1 U DNase (Invitrogen) to generate fragments of 50−100 bp range followed by heat inactivation by incubation for 10 minutes at 95°C. Fragmented DNA was labeled with 5.1 nmol each of biotin-16-ddUTP and biotin-16-dUTP (Roche) using 1360 units of recombinant terminal deoxynucleotidyl transferase enzyme (Roche at 400 U/µl) in a 75 µl reaction in the presence of 1× one-phor-all buffer (Amersham), by incubation at 37°C for 90 minutes followed by heat-inactivation for 10 minutes at 99°C. The labeled DNA sample was purified using a 96-well 3K plate (Pall Scientific) by addition of 170 µl of water to the labeling reaction prior to loading a single well per reaction. The 3K plate was fitted onto a vacuum manifold with a pressure (25−30 in. Hg) for 2−3 hours or until samples appeared visibly dry. The labeled purified DNA sample was eluted from the 3K

filter well by placing 56 µL of water on the filter surface followed by a gentle vortex of the entire plate for 15 minutes.

Signal Detection

The purified labeled DNA was combined with non-specific DNA carriers (1µl of Cot-1 @ 10ug/ul, 8µl of HS-DNA @ 10µl/µl, 8µl of yeast tRNA @ 10µg/µl) and denatured for 10 minutes at 95C. After denaturation, 139µl of hybridization buffer was added to yield the final concentrations of 10mM Tris ph 8, 3M TMACL, 0.1% Tx-100 and the repetitive sequences were pre-blocked by a 60 minute hybridization for 1 hour at 50C. Subsequent to this pre-blocking step, formamide was added to a final concentration of 3% and this mixture was then hybridized to the high-density oligonucleotide array at 50°C for 12−16 hours. All signal detection steps were performed using an in house built fluidics station to allow parallel processing of 192 arrays.

The arrays were washed in 6× SSPE buffer briefly and subjected to a low salt stringency wash by incubation in 0.2× SSPE for 30 minutes at 42°C followed by a brief rinse in MES buffer. For signal detection, the arrays were incubated with 5µg/ml streptavidin (Sigma Aldrich) for 15 minutes at 25°C, followed by 1.25 ug/ml biotinylated anti-streptavidin antibody (Vector Laboratories) for 10 minutes at 25°C, then 1 ug/ml streptavidin-Cy-chrome conjugate (Molecular Probes) for 15 minutes at 25°C. The 1.25ug/ml biotinylated anti-streptavidin antibody step followed by the 1ug/ml streptavidin-Cy-chrome conjugate step was repeated for signal amplification. The arrays were then subjected to low salt stringency wash by incubation in 0.2× SSPE buffer for 30 minutes at 45°C. The hybridization of labeled DNA was detected by measuring Cy-chrome fluorescence using a custom built confocal laser scanner (Perlegen Sciences).

Supplementary Table 1 Summary of genotype discordance by amplicon quality score

Amplicon Quality Score	Mean percent discordance (Affy 500K)	Mean percent discordance (other platforms)	Percent SNPs in consensus
0.05	12.2	16.8	0.3
0.15	4.6	5.9	0.8
0.25	1.6	2.9	1.4
0.35	1.1	1.4	2.5
0.45	0.7	1.0	4.7
0.55	0.4	0.6	7.3
0.65	0.3	0.5	18.0
0.75	0.2	0.5	23.2
0.85	0.2	0.4	23.4
0.95	0.2	0.4	18.3

Supplementary Table 2: Summary of SNPs with discrepant genotypes in individuals genotyped by both HapMap and Seattle SNPs

HapMap Phase II	Seattle SNPs	Number of SNPs
Homozygous	Homozygous	8
Homozygous	Heterozygous	57
Heterozygous	Homozygous	45
Heterozygous	Heterozygous	1
Other		3

Supplementary Table 3. Summary of genotype discrepancies identified by comparison with fosmid-end sequencing

Center	Platform	Number of genotype calls analysed	Percent discrepant
illumina	Illumina - Infinium	290,536	0.06%
imsut-riken	Invader	114,615	0.22%
illumina	Illumina – GoldenGate	139,698	0.32%
mcgill-gqic	Illumina – GoldenGate	57,939	0.46%
chmc	Illumina – GoldenGate	39,025	0.46%
ucsf-wu	FP-TDI	7,074	0.85%
sanger	Illumina – GoldenGate	120,955	0.86%
bcm	MIP	29,990	0.95%
broad	Sequenom	13,436	1.24%
perlegen	Perlegen	1,018,457	1.43%
broad	Illumina – GoldenGate	30,396	1.59%
chmc	Sequenom	11,667	1.64%

Supplementary Table 4. Summary of monomorphism/polymorphism discrepancies by genotyping platform compared to Infinium platform

Platform	No. comparisons[1]	Mono/Poly discrepancy rate	Ratio 'other':Infinium in calling monomorphic	Percent MAF<0.1
Affymetrix	203,196	0.08%	0.4	99
Illumina: GoldenGate	258,520	0.07%	1.7	75
Invader	114,081	0.07%	3.7	69
Perlegen	108,507	0.42%	3.5	56
Illumina: Infinium	257,164	0.001%	NA	100

Excludes FP-TDI and MIP platforms due to insufficient data

Supplementary Table 5. Summary of genotype submissions in Phase II HapMap (Release 21)

		YRI			CEU			CHB+JPT		
Phase	Center	QC+	QC-	Total	QC+	QC-	Total	QC+	QC-	Total
I	Affymetrix				112,046	379	112,425			
	BCM	52,989	2,047	55,036	53,763	4,186	57,949	51,060	3,295	54,355
	Broad	196,790	19,887	216,677	91,981	11,622	103,603	198,717	17,671	216,388
	CHMC	90,616	12,784	103,400	95,790	17,033	112,823	92,503	11,248	103,751
	Illumina	260,699	34,736	295,435	260,529	27,338	287,867	261,159	34,296	295,455
	RIKEN	203,388	20,387	223,775	220,850	29,464	250,314	210,343	16,157	226,500
	McGill-GQIC	99,688	15,220	114,908	104,,680	12,221	116,901	99,657	15,238	114,895
	Perlegen				5,494	14	5,508			
	Sanger	234,971	20,976	255,947	231,548	22,658	254,206	236,191	19,577	255,768
	UCSF-WU	11,419	808	12,227	14,438	1,788	16,226	11,298	790	12,088
Total		1,150,560	126,845	1,277,405	1,191,119	126,703	1,317,822	1,160,928	118,272	1,279,200
II	Affymetrix	489,925	3,468	493,393	490,789	2,604	493,393	491,266	2,258	493,524
	Perlegen	2,687,260	1,891,130	4,578,390	2,740,703	1,837,694	4,578,397	2,780,503	1,796,673	4,577,176
Total		3,177,185	1,894,598	5,071,783	3,231,492	1,840,298	5,071,790	3,271,769	1,798,931	5,070,700
Overall		4,327,745	2,021,443	6,349,188	4,422,611	1,967,001	6,389,612	4,432,697	1,917,203	6,349,900

Supplementary Table 6. Pairwise identity-by-state (IBS) sharing between and within subpopulation..

Mean (SD) N min – max	YRI	CEU	CHB	JPT
	0.819 (0.00067)			
YRI	3933			
	0.816 – 0.821			
	0.779 (0.00074)	0.837 (0.0009)		
CEU	8100	3940		
	0.775 – 0.781	0.833 – 0.841		
	0.778 (0.00078)	0.814 (0.00091)	0.850 (0.00095)	
CHB	4050	4050	990	
	0.774 – 0.781	0.812 – 0.817	0.847 – 0.854	
	0.778 (0.00087)	0.814 (0.0009)	0.849 (0.00095)	0.851 (0.0011)
JPT	3960	3960	1980	946
	0.773 – 0.781	0.810 – 0.817	0.845 – 0.852	0.846 – 0.854

Supplementary Table 7. Subjects with multiple non-adjoining homozygous regions > 3 Mb.

Subj. ID	Chrom.	Start pos.	End pos.	Length (bp)	SNP ct.	Het. Ct.
YRI						
na18502	3	47,546,444	51,820,266	4,273,822	1,800	9
na18502	3	129,667,114	132,912,728	3,245,614	2,444	39
na18855	6	58,260,082	63,509,818	5,249,736	1,976	12
na18855	3	82,649,575	86,219,997	3,570,422	3,085	10
na19093	14	75,643,163	81,851,230	6,208,067	7,957	16
na19093	23	104,208,008	108,837,101	4,629,093	2,066	6
na19172	1	211,515,431	220,659,004	9,143,573	11,054	25
na19172	6	57,237,646	65,457,337	8,219,691	4,479	22
na19172	10	36,923,959	44,771,184	7,847,225	5,367	34
CEU						
na10847	23	104,226,505	108,664,817	4,438,312	1,987	21
na10847	23	55,226,270	58,305,966	3,079,696	1,166	29
na11993	11	46,634,310	56,382,761	9,748,451	4,837	17
na11993	11	64,059,102	67,060,543	3,001,441	1,513	26
na12740	19	19,997,049	33,628,437	13,631,388	4,400	10
na12740	16	68,073,669	71,572,247	3,498,578	2,626	18
na12874	1	145,991,559	239,297,570	93,306,011	103,773	247
na12874	6	46,264,500	50,535,986	4,271,486	5,207	9
na12892	20	24,728,544	29,962,987	5,234,443	1,703	35
na12892	3	50,344,550	53,671,328	3,326,778	2,082	54
CHB						
na18537	11	50,256,797	56,314,992	6,058,195	2,536	8
na18537	10	36,687,723	41,825,614	5,137,891	2,108	11
JPT						
na18981	6	58,878,583	63,922,941	5,044,358	2,048	16
na18981	8	51,624,066	56,193,067	4,569,001	5,225	7
na18987	14	33,695,888	73,065,210	39,369,322	44,249	40
na18987	18	7,092,706	26,082,693	18,989,987	19,780	57
na18987	4	77,472,732	88,862,875	11,390,143	11,325	31
na18987	6	37,447,729	47,883,332	10,435,603	12,696	28
na18987	7	82,014,909	92,178,675	10,163,766	11,542	18
na18987	8	111,512,918	121,252,676	9,739,758	12,235	20
na18987	8	72,562,269	80,014,969	7,452,700	8,736	7
na18987	9	96,082,819	103,020,889	6,938,070	8,090	16
na18987	6	25,792,585	29,392,330	3,599,745	3,699	15
na18992	3	72,306,277	115,134,904	42,828,627	38,231	30
na18992	6	55,749,642	80,761,476	25,011,834	26,231	56
na18992	2	17,911,128	42,249,929	24,338,801	30,249	32
na18992	13	71,814,922	94,623,392	22,808,470	29,846	57
na18992	4	6,725,243	26,274,662	19,549,419	24,256	23
na18992	16	24,045	19,196,013	19,171,968	22,505	23
na18992	3	46,668,436	51,329,728	4,661,292	2,072	14
na18992	2	94,794,129	98,098,204	3,304,075	1,073	5
na18992	2	237,204,845	240,369,042	3,164,197	3,653	11

Supplementary Table 8. Subjects with single homozygous regions > 3 Mb

YRI

Subj. ID	Chrom.	Start pos.	End pos.	Total length (bp)	SNP ct.	Het. Ct.
na19201	5	65,630,035	88,242,996	22,612,961	21,208	23
na18501	2	129,153,735	143,724,758	14,571,023	16,660	29
na19140	18	23,614,010	34,759,766	11,145,756	14,069	29
na18503	4	73,758,178	83,381,108	9,622,930	9,776	37
na19211	3	90,355,175	98,847,728	8,492,553	3,467	4
na19171	4	101,029,288	109,492,748	8,463,460	7,553	13
na19161	20	24,728,544	31,060,133	6,331,589	2,450	15
na19206	4	47,375,945	52,759,666	5,383,721	1,419	10
na19153	19	19,077,897	24,216,651	5,138,754	3,793	13
na18506	15	38,381,231	42,883,194	4,501,963	3,009	10
na18508	17	28,829,743	32,500,734	3,670,991	3,942	16
na19205	20	33,318,157	36,984,349	3,666,192	3,146	9
na19154	3	95,392,747	98,838,920	3,446,173	3,345	4
na19092	9	68,238,389	71,548,897	3,310,508	4,169	10
na19101	8	112,950,254	116,179,851	3,229,597	3,581	14
na19138	6	65,518,992	68,744,069	3,225,077	4,817	10
na18870	1	39,822,234	43,012,800	3,190,566	3,069	14
na19141	3	82,865,889	86,011,270	3,145,381	2,693	10

CEU

na07056	2	192,353,417	199,819,815	7,466,398	7,593	22
na10855	3	88,542,644	95,479,875	6,937,231	1,518	20
na12864	5	44,466,290	50,204,271	5,737,981	1,365	11
na12003	6	25,788,389	31,087,063	5,298,674	7,553	18
na10838	10	37,838,976	41,735,506	3,896,530	932	16
na12249	6	58,878,583	62,755,705	3,877,122	790	3
na06993	3	163,082,396	166,395,422	3,313,026	3,603	12
na10846	5	128,765,181	131,921,228	3,156,047	2,679	15

CHB

na18612	11	44,957,731	61,824,394	16,866,663	11,954	50
na18529	8	42,569,476	49,396,402	6,826,926	1,502	17
na18632	5	103,103,615	106,968,418	3,864,803	4,065	11
na18623	10	73,535,911	76,631,183	3,095,272	1,894	15
na18558	5	39,603,972	42,661,512	3,057,540	3,744	27

JPT

na18964	3	78,408,132	87,126,013	8,717,881	7,797	14
na18994	3	78,665,898	86,114,668	7,448,770	6,489	10
na18967	8	42,773,387	49,706,945	6,933,558	1,560	13
na18976	6	27,798,432	33,582,400	5,783,968	10,525	29
na18972	11	50,669,978	56,097,300	5,427,322	2,131	17
na18974	2	185,100,373	190,272,687	5,172,314	5,370	6
na18975	12	82,854,504	86,272,156	3,417,652	3,421	24

Supplementary Table 9. Candidate regions for recent adaptive evolution by LRH and iHS tests

Chr	Bin start	Bin end	Test	Population	Genes in region	Peak SNP
1	35100000	35300000	IHS	CEU	ZMYM6, ZMYM1	rs11263952
1	65800000	65900000	LRH	CHB+JPT	LEPR	rs4655795
1	68400000	68550000	LRH	CEU	GPR177	rs7516564
1	70200000	70400000	LRH	YRI	LRRC7, LRRC40, SFRS11	rs7518536
1	73050000	73650000	LRH	CHB+JPT		rs12567259
1	76200000	76300000	IHS	CHB+JPT	ST6GALNAC3	rs12040836
1	82800000	82950000	IHS	CHB+JPT		rs9324198
1	90450000	90600000	LRH	CEU		rs7528896
1	92850000	93050000	IHS	CHB+JPT	EVI5, RPL5, FAM69A	rs1337107
1	94050000	94150000	LRH	CHB+JPT	DNTTIP2, GCLM	rs10874811
1	106350000	106500000	LRH	YRI		rs11184772
1	157359782	157359782	IHS	CEU	CD84	rs2369722
1	157850000	157950000	IHS	YRI	ARHGAP30, PVRL4, KARCA1, PFDN2, NIT1, DEDD, UFC1, USP21, PPOX	rs11265554
1	165850000	166100000	Both	CHB+JPT	NME7, BLZF1, C1orf114	rs2300158
1	167900242	167900242	IHS	CEU	FMO2	rs2020862
1	169450000	169550000	IHS	CHB+JPT		rs4916195
1	186500000	186650000	IHS	CEU		rs12066792
1	193450000	193550000	LRH	YRI	CFHR3, CFHR1	rs644598
1	216200000	216300000	LRH	CEU		rs1415995
1	219650000	219750000	IHS	YRI		rs17661703
2	7900000	8050000	LRH	CEU		rs976036
2	9700000	9800000	LRH	CHB+JPT	YWHAQ	rs7424240
2	21650000	21750000	IHS	YRI		rs10197373
2	24650000	24850000	LRH	YRI	NCOA1	rs995648
2	73800000	73950000	IHS	CEU	LOC200420, CML2, TPRKB, DUSP11	rs12998980
2	83300000	83550000	LRH	CHB+JPT		rs11693198
2	89300000	89450000	IHS	CEU	LOC651928	rs1874935
2	108250000	109100000	IHS	CHB+JPT	SULT1C3, SULT1C1, SULT1C2, GCC2, FLJ38668, LIMS1, RANBP2, FLJ32745, EDAR	rs10175540
2	121550000	121700000	LRH	CEU	TFCP2L1	rs6723834
2	135000000	136550000	Both	CEU	MGAT5, TMEM163, ACMSD, CCNT2, YSK4, RAB3GAP1, ZRANB3, R3HDM1, UBXD2, LCT, MCM6, DARS	rs1446584
2	137000000	137250000	LRH	CEU		rs12691894
2	157950000	158050000	IHS	CEU	GALNT5, KIAA1189	rs3214040
2	159100000	159250000	LRH	CHB+JPT	LOC130940, PKP4	rs1117199
2	178250000	178450000	Both	CEU	TTC30A, PDE11A	rs4407279
2	192950000	193050000	IHS	YRI		rs1596880
2	194650000	194900000	IHS	YRI		rs6710933
2	197200000	197300000	IHS	CHB+JPT	HECW2	rs6719725
2	226450000	226600000	LRH	CEU		rs873024
3	17450000	17550000	IHS	CHB+JPT	TBC1D5	rs7650295
3	25800000	26300000	IHS	CEU/CHB+JPT	OXSM	rs4681035
3	36150000	36250000	LRH	CEU		rs11720944
3	49300000	49650000	IHS	CHB+JPT	USP4, GPX1, RHOA, TCTA, AMT, NICN1, DAG1, BSN	rs7622302
3	56550000	56700000	IHS	YRI	CCDC66, C3orf63	rs282533

3	72650000	72750000	IHS	CEU		rs13066103
3	79150000	79250000	LRH	YRI	ROBO1	rs4234349
3	87300000	87400000	LRH	YRI	CHMP2B, POU1F1	rs12635997
3	90150000	90300000	LRH	CHB+JPT		rs6551450
3	106100000	106250000	IHS	CHB+JPT		rs9846552
3	127050000	127150000	LRH	CHB+JPT	LOC200810	rs4679199
3	134450000	134550000	IHS	YRI	TMEM108	rs4854579
3	140600000	140700000	IHS	CHB+JPT	RBP2	rs12695698
3	146750000	146900000	LRH	YRI		rs2375839
3	162300000	162400000	LRH	CHB+JPT	B3GALNT1	rs4618258
3	165250000	165400000	LRH	YRI		rs1449936
3	189650000	189800000	IHS	CHB+JPT	LPP	rs1019673
3	197000000	197150000	LRH	YRI	MUC4, TNK2	rs7636635
4	20650000	20950000	LRH	YRI	KCNIP4	rs6854888
4	33600000	34700000	Both	CEU/CHB+JPT/YRI		rs11934714
4	41300000	41400000	IHS	CEU/CHB+JPT	DKFZP686A01247	rs4343753
4	41900000	42050000	IHS	CHB+JPT	SLC30A9, CCDC4	rs2343617
4	56100000	56250000	IHS	YRI	TMEM165, CLOCK	rs9312661
4	85700000	85850000	LRH	CEU	NKX6-1	rs1444961
4	93850000	94050000	LRH	CHB+JPT	GRID2	rs970405
4	100000000	101000000	LRH	CHB+JPT	EIF4E, METAP1, ADH5, ADH4, ADH6, ADH1A, ADH1B, ADH1C, ADH7, C4orf17, RG9MTD2, MTTP	rs1348276
4	104750000	104900000	LRH	CEU	TACR3	rs2903341
4	123550000	123650000	IHS	YRI		rs13114649
4	132900000	133000000	IHS	CEU		rs7687345
4	144100000	144550000	Both	CHB+JPT	USP38	rs877032
4	145300000	145400000	LRH	CEU	GYPA	rs7657795
4	148450000	148600000	IHS	YRI		rs1354886
4	158900000	159100000	IHS	CHB+JPT		rs11934695
4	163950000	164100000	LRH	CHB+JPT		rs1003527
4	171800000	171950000	LRH	CHB+JPT		rs444538
4	176600000	176750000	IHS	CEU		rs7653918
4	190900000	191050000	LRH	CHB+JPT		rs6820482
5	24300000	24550000	LRH	CEU	CDH10	rs1346511
5	64850000	65100000	Both	CHB+JPT	CENPK, PPWD1, TRIM23, FLJ13611, LOC643079, SGTB, NLN	rs3855589
5	110150000	110300000	LRH	CEU		rs6594483
5	112350000	112550000	LRH	CEU	DCP2, MCC	rs9326874
5	120550000	120950000	LRH	CHB+JPT		rs2406518
5	170400000	170500000	LRH	CHB+JPT	RANBP17	rs10070298
6	18700000	18850000	LRH	CHB+JPT		rs6459629
6	33550000	33700000	LRH	YRI	BAK1, FLJ43752, ITPR3	rs210209
6	47350000	47850000	LRH	CHB+JPT	TNFRSF21, CD2AP, GPR111, GPR115	rs1032146
6	48300000	48400000	LRH	CHB+JPT		rs325049
6	63500000	63650000	LRH	YRI		rs6453796
6	70100000	70250000	LRH	YRI	BAI3	rs6939864
6	74950000	75050000	LRH	YRI		rs9359077
6	77900000	78000000	LRH	YRI		rs9359255
6	81800000	81950000	LRH	CEU		rs9359454
6	83400000	83850000	LRH	CHB+JPT	C6orf157, DOPEY1	rs1547251
6	84800000	85000000	IHS	CEU	C6orf117, KIAA1009	rs9449802

6	122800000	122950000	IHS	YRI	SERINC1, PKIB	rs10080477
6	125950000	126100000	LRH	CHB+JPT		rs2211418
6	130550000	130650000	IHS	YRI	SAMD3	rs9483097
7	20100000	20250000	IHS	YRI	ITGB8	rs3757727
7	73450000	74750000	IHS	CHB+JPT	LOC442582, GTF2IRD1, GTF2I, GTF2IRD2, PMS2L5, WBSCR16, GTF2IRD2B, NCF1, LOC441257, PMS2L2, DKFZP434A0131, LOC442578, LOC541473, TRIM74, TRIM73, NSUN5B	rs2527366
7	74817831	74817831	IHS	CEU	PMS2L3, HIP1	rs1167796
7	88000000	88100000	LRH	YRI	FLJ32110, MGC26647	rs10229796
7	104450000	104550000	IHS	CEU	SRPK2	rs12538590
7	105600000	105750000	LRH	CHB+JPT		rs6466108
7	111750000	111950000	Both	CHB+JPT		rs4473967
7	124100000	124250000	IHS	CEU	POT1, LOC401398	rs4463363
7	141500000	142150000	Both	CHB+JPT/YRI	LOC647353, PRSS1, PRSS2, EPHB6, TRPV6, TRPV5	rs2855918
8	9500000	9900000	Both	CHB+JPT/YRI	TNKS	rs6994574
8	11200000	11300000	IHS	CHB+JPT	MTMR9, AMAC1L2	rs6991606
8	50300000	50400000	IHS	YRI		rs3925383
8	51050000	52150000	LRH	CEU/CHB+JPT	SNTG1	rs6473486
8	52600000	53050000	LRH	CEU	PCMTD1	rs16916598
8	111900000	112050000	LRH	CEU		rs10808439
9	11800000	11900000	LRH	YRI		rs10809610
9	12600000	12700000	IHS	CEU	TYRP1	rs10960749
9	24350000	24450000	IHS	YRI		rs12339773
9	42850000	44200000	IHS	CEU/CHB+JPT/YRI		rs4929025
9	64250000	64450000	IHS	CEU		rs11262451
9	68050000	68250000	IHS	CHB+JPT	CBWD3, FOXD4L2, FOXD4L3, PGM5	rs12554575
9	87900000	88050000	LRH	CEU		rs10512193
9	97700000	97850000	IHS	YRI	C9orf156, HEMGN, ANP32B	rs3780419
9	103900000	104000000	LRH	CHB+JPT	SMC2	rs4742902
9	108250000	108400000	IHS	CHB+JPT		rs10121673
9	127900000	128200000	Both	CEU/CHB+JPT	C9orf90, SLC25A25, PTGES2, LOC389791, LCN2, C9orf16, CIZ1, DNM1, GOLGA2, TRUB2, COQ4, SLC27A4	rs6478813
9	137000000	137150000	IHS	YRI	C9orf86, PHPT1, MAMDC4, EDF1, TRAF2, FBXW5, C8G, LCN12, PTGDS	rs2784075
10	2950000	3100000	IHS	CEU/CHB+JPT	PFKP	rs10903912
10	11000000	11150000	LRH	YRI	CUGBP2	rs201093
10	55600000	55750000	IHS	CHB+JPT	PCDH15	rs7915662
10	60700000	60850000	LRH	CEU	FAM13C1	rs284643
10	84000000	84100000	LRH	CEU	NRG3	rs1414772
10	94950000	95050000	IHS	CHB+JPT		rs7091432
10	102200000	102400000	IHS	YRI	WNT8B, SEC31L2, NDUFB8, HIF1AN	rs9420797
10	107250000	107350000	IHS	CHB+JPT		rs4918165
10	109650000	109800000	IHS	CHB+JPT		rs2151876
11	5116672	5116672	LRH	YRI	OR52A4, OR52A5, OR52A1, HBB	rs2472528
11	10650000	10800000	LRH	CHB+JPT	MRVI1, CTR9, EIF4G2	rs10840479
11	25250000	25600000	Both	CHB+JPT		rs2404085
11	34900000	35050000	LRH	CEU	PDHX	rs2732564

11	38400000	38750000	LRH	CEU/YRI		rs11034801
11	48450000	48950000	Both	CHB+JPT	OR4A47	rs2865636
11	61300000	61450000	LRH	CHB+JPT	C11orf9, C11orf10, FEN1, FADS1, FADS2, FADS3, RAB3IL1	rs2072114
11	63450000	63550000	IHS	CHB+JPT	NAT11, COX8A, OTUB1, LRP16	rs539432
11	81300000	81750000	LRH	CHB+JPT		rs605296
11	119550000	119700000	LRH	CEU	OAF, POU2F3	rs11217785
12	2850000	2950000	IHS	CEU	FOXM1, C12orf32, TULP3, TEAD4	rs10774069
12	18400000	18500000	IHS	YRI	PIK3C2G	rs11044109
12	21800000	21900000	LRH	YRI	KCNJ8, ABCC9	rs1283822
12	30400000	30500000	LRH	YRI		rs11050884
12	34550000	36150000	LRH	YRI		rs11829528
12	39800000	39950000	LRH	CHB+JPT		rs4768334
12	45350000	45500000	LRH	YRI	SLC38A4	rs2408619
12	75300000	75400000	IHS	CEU	OSBPL8	rs12826628
12	78000000	78650000	Both	YRI	SYT1, PAWR	rs7955388
12	109750000	109950000	LRH	CEU	CCDC63, MYL2, CUTL2	rs4766517
12	125600000	125750000	LRH	CEU		rs1205378
13	24250000	24350000	IHS	CHB+JPT	RNF17	rs2305369
13	56500000	57100000	LRH	YRI	FLJ40296	rs473750
13	61100000	61350000	LRH	CHB+JPT		rs4884396
13	62700000	62850000	Both	CHB+JPT		rs9564023
13	67150000	67350000	LRH	YRI		rs1411886
13	75100000	75250000	IHS	YRI	LMO7	rs9318370
14	19449360	19489709	LRH	CEU/YRI	OR4K5, OR4K1	rs1780906
14	27550000	28050000	Both	CHB+JPT		rs1958743
14	47700000	47850000	LRH	YRI		rs10141880
14	69950000	70050000	LRH	CEU	SYNJ2BP, ADAM21	rs12889741
14	105800000	105900000	LRH	YRI	IGHG1	rs4774094
15	43000000	43150000	IHS	CEU	C15orf43, SORD	rs414966
15	53250000	53700000	LRH	YRI	C15orf15, RAB27A, PIGB, CCPG1, DYX1C1, PYGO1, PRTG	rs16953251
15	62150000	62300000	IHS	CHB+JPT	FAM96A, SNX1, SNX22, PPIB, CSNK1G1	rs3816385
15	64000000	64100000	LRH	CHB+JPT	MEGF11	rs441949
15	75550000	75650000	LRH	YRI	HMG20A	rs12917044
16	1450000	1600000	LRH	CEU	CLCN7, LOC390667, KIAA0683, IFT140, C16orf30	rs2064289
16	14450000	14550000	IHS	YRI	PARN	rs7184698
16	17300000	17450000	IHS	CHB+JPT	XYLT1	rs7500021
16	22850000	22950000	IHS	YRI		rs12919791
16	31400000	31950000	IHS	CEU/YRI	SLC5A2, C16orf58, ERAF, ZNF720, ZNF267	rs2136013
16	34050000	45100000	IHS	CEU/CHB+JPT/YRI	FLJ43980	rs4887582
16	64200000	64350000	IHS	CHB+JPT		rs8057899
16	74100000	74550000	Both	CHB+JPT/YRI	CHST5, GABARAPL2, ADAT1, KARS, TERF2IP	rs8061878
16	78350000	78450000	IHS	CEU		rs7205712
17	18400000	18500000	LRH	YRI	FLJ36492, FLJ40244	rs6502661
17	56150000	56450000	IHS	CHB+JPT	BCAS3	rs747895
17	61750000	61850000	IHS	CEU	PRKCA	rs8075066
18	7500000	7650000	IHS	CEU	PTPRM	rs489659
18	14600000	15150000	Both	CEU/CHB+JPT	ANKRD30B	rs1811759
18	28800000	29200000	LRH	YRI	C18orf34	rs443593
18	38800000	39250000	LRH	CEU	RIT2, SYT4	rs879215

18	68900000	69050000	IHS	CEU		rs10871712
18	70800000	70900000	IHS	CEU		rs12971033
19	43400000	43600000	Both	YRI	DPF1, PPP1R14A, SPINT2, LOC541469, C19orf33, YIF1B, KCNK6, C19orf15, PSMD8, GGN, SPRED3, FAM98C, RASGRP4	rs4312417
19	45200000	45300000	IHS	CEU	ZNF546, LOC163131, LOC284323	rs234352
20	6850000	7000000	LRH	CEU		rs6140141
20	33700000	33900000	LRH	YRI	CPNE1, RBM12, NFS1, C20orf52, RBM39, PHF20	rs2425090
20	35850000	35950000	IHS	YRI	CTNNBL1	rs2294441
20	36750000	36950000	IHS	YRI	SLC32A1, ACTR5, PPP1R16B	rs6129111
22	29250000	29500000	LRH	YRI	GAL3ST1, PES1, TCN2, SLC35E4, DUSP18, OSBP2	rs4820888
22	32350000	32650000	LRH	YRI	LARGE	rs2267267
22	34800000	35100000	LRH	YRI	APOL3, APOL4, APOL2, APOL1, MYH9	rs132683
22	45650000	45800000	LRH	CHB+JPT	TBC1D22A	rs1807721
X	18850000	19050000	IHS	CEU	GPR64	rs5955721
X	26600000	26700000	IHS	CHB+JPT		rs1842186
X	30150000	30300000	IHS	CHB+JPT		rs2867195
X	32300000	32400000	LRH	YRI	DMD	rs808540
X	34900000	35350000	Both	YRI		rs16991838
X	41150000	41300000	IHS	YRI	CASK	rs13440974
X	57700000	61850000	IHS	CEU/CHB+JPT	ZXDA	rs7392401
X	61800000	65200000	IHS	CEU/CHB+JPT	LOC139886, ARHGEF9, FLJ39827, ASB12, MTMR8, KIAA1166, ZC3H12B, LAS1L, MSN, VSIG4, HEPH	rs12388294
X	66200000	66500000	IHS	CHB+JPT		rs12556495
X	72450000	72550000	IHS	CHB+JPT	CDX4	rs4892781
X	87250000	87400000	IHS	CHB+JPT		rs5924296
X	88050000	88300000	LRH	CHB+JPT		rs5942366
X	98400000	99000000	IHS	CEU/YRI		rs1832648
X	121100000	121200000	IHS	CHB+JPT		rs2495677
X	134550000	134700000	IHS	YRI	CT45-1, CT45-2, CT45-4, CT45-3, CT45-5, CT45-6	rs2254857
X	146800000	147000000	LRH	CEU	FMR1NB	rs6525878
X	154200000	154500000	IHS	CHB+JPT	F8A1, F8A2, F8A3, H2AFB1, H2AFB3, H2AFB2, TMLHE	rs622581

Reference List

1. Przeworski, M., Hudson, R. R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet.* **16**, 296-302 (2000).

2. Frisse, L. *et al.* Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**, 831-843 (2001).

3. Fu, Y. X. Statistical properties of segregating sites. *Theor. Popul. Biol.* **48**, 172-197 (1995).

4. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320 (2005).

5. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725-1729 (2001).

6. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

7. Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. & Hill, W. G. Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* **15**, 1468-1476 (2005).

8. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

9. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).

10. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072-1079 (2005).

11. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005-1017 (2001).

12. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat. Genet.* **39**, 906-913 (2007).

13. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-668 (2007).

14. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-324 (2005).

15. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581-584 (2004).

Legends to supplementary figures

Supplementary Figure 1. Characteristics of Perlegen amplicons and matched non-Perlegen and Perlegen Phase II assays.

A) Frequency histogram of amplicon length for amplicon primer pairs that mapped uniquely to NCBI Human genome Build 35.1 (n = 296,273). B) Frequency distribution of the number of SNPs on each amplicon (amp.ct. = 301,944; SNP ct. = 4,420,481). C) Frequency histogram of amplicon quality score.

Supplementary Figure 2. Amplicon quality score analysis of genotype discordance between non-Perlegen and Perlegen Phase II assays.

303,660 SNPs were selected from the redundant/unfiltered dataset that were QC+ across all three analysis panels for both a non-Perlegen (NPRL) and a Perlegen Phase II (PRL) assay. Each NPRL/PRL data pair was binned based on the amplicon quality score (AQS) of the corresponding amplicon for the PRL SNP, and the proportion of discordant genotypes was calculated for each data pair. A) The proportion of discordant genotypes was plotted against the reference allele frequency from the non-Perlegen assay. Points with discordance > 0.01 and allele frequency between 0.02 and 0.98 were plotted in red (high discordants), while other points were plotted in blue (low discordants). B) The reference allele frequency from PRL plotted against that from NPRL. Density was estimated individually for each plot, but red/blue color assignment was based on the filter described in A. For better frequency visualization, a random thinning algorithm was used to equalize the number of plotted points to that of the lowest AQS bin. (AQS 0-0.2; n=5,289). The dataset used in this figure came from the redundant/unfiltered dataset from release 21.

Supplementary Figure 3. Patterns of inter-chromosomal LD.

For each analysis panel we identified common (MAF≥0.05) SNPs that show strong association with a SNP on another chromosome. These are classified into those that show no strong association to other SNPs near to the catalogued location and which are therefore most likely the result of mis-mapping (grey lines) and those that show strong inter and intra-chromosomal association (red lines). Also shown is the location of segmental duplications[11] (yellow bars). A cut-off on the likelihood ratio test statistic for association was used to identify

SNP pairs. The apparent larger number of SNPs showing inter-chromosomal LD in the CHB+JPT panel simply reflects the larger sample size.

Supplementary Figure 4. Comparison of Phase I and Phase II HapMap

Features of A) SNP spacing, B) the decay of LD with distance, C) minor allele frequency and D) derived allele frequency in the Phase I and Phase II HapMap data.

Supplementary Figure 5. Model based imputation of genotypes from tagging SNPs.

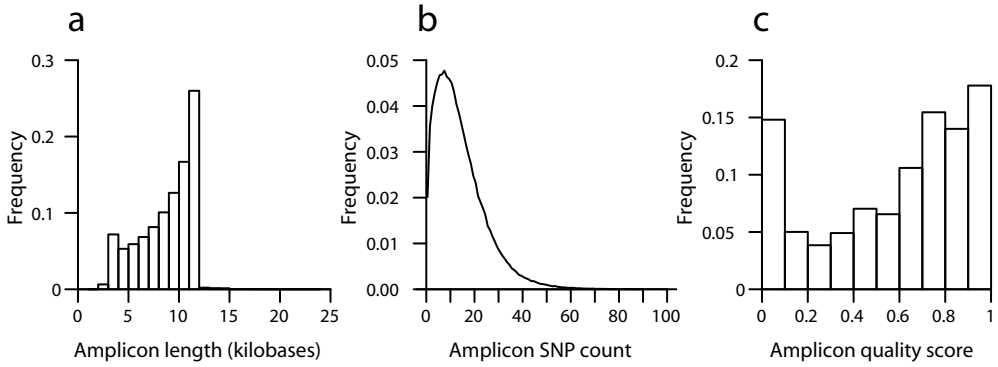
For the HapMap-ENCODE region ENr321 on 8q24.11 in YRI we used recently developed statistical methodology[12,13] to impute genotypes using SNPs present on the Affymetrix GeneChip 500K as tags. Briefly, for each of the 120 parents we imputed genotypes at Phase II HapMap SNPs not present on the array using phased haplotypes at all Phase II SNPs from the other 119 individuals. For each imputed SNP with MAF>0.2 we calculate the square of the correlation between expected genotype value (coded as 0, 1 and 2) and the observed genotype value (red circles). For the same SNPs we also calculate the maximum r^2 to any of the array SNPs within the region (black crosses). Because the imputation methodology requires an estimate of the fine-scale structure of recombination rate variation, the recombination rate estimated from Phase II HapMap is also shown. Across the region the average imputation r^2 is 0.86 compared to an average max r^2 of 0.59. Regions of low imputation success typically correspond to regions of low SNP density and high recombination rate.

Supplementary Figure 6. The distribution of recombination for each chromosome.

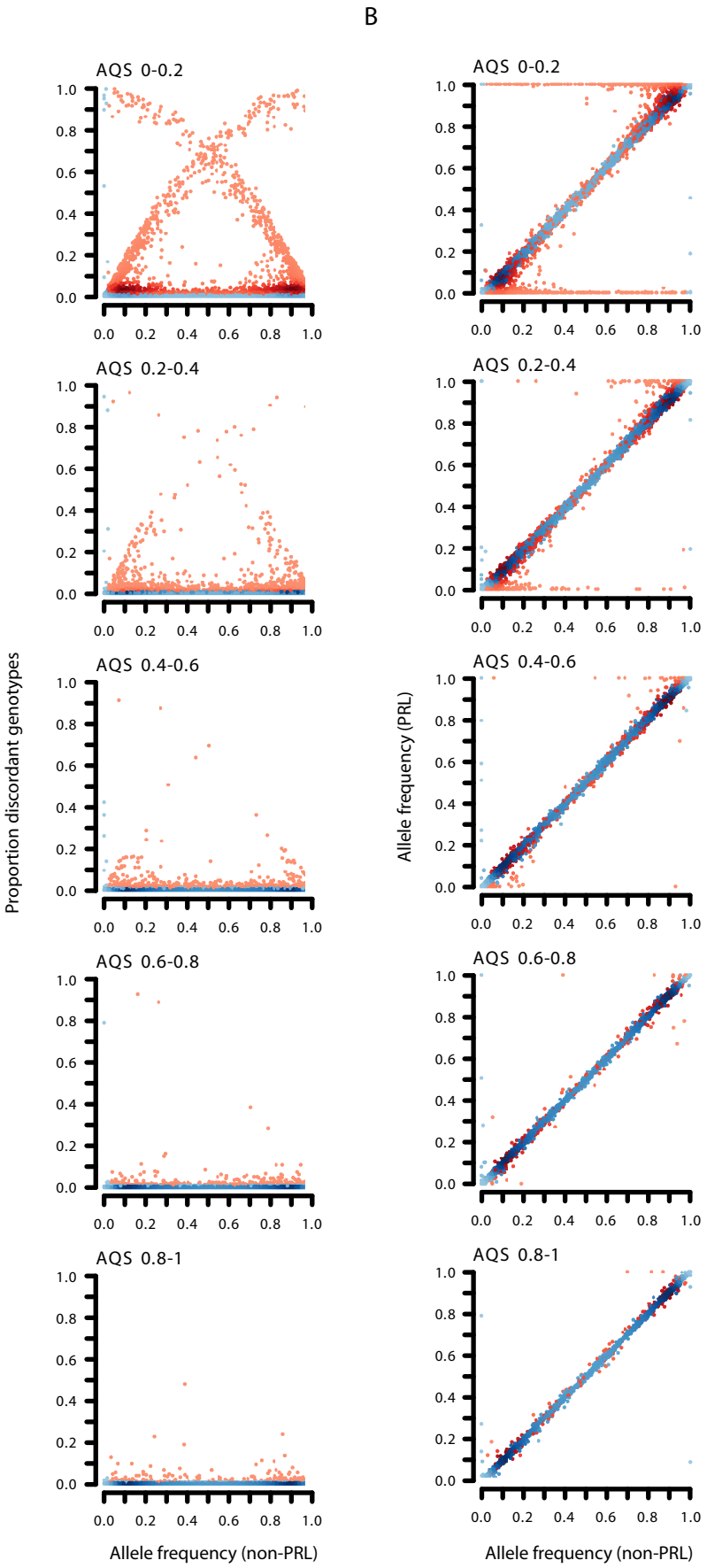
Each curve shows the concentration of recombination into recombination hotspots[14,15]. For each chromosome SNP intervals are ordered by estimated genetic map length (starting with the highest). The proportional summed genetic map length is plotted against the proportional summed physical distance. If recombination rate were uniform we would observe a straight line.

Supplementary Figure 7. Gene ontology and recombination hotspot motif density

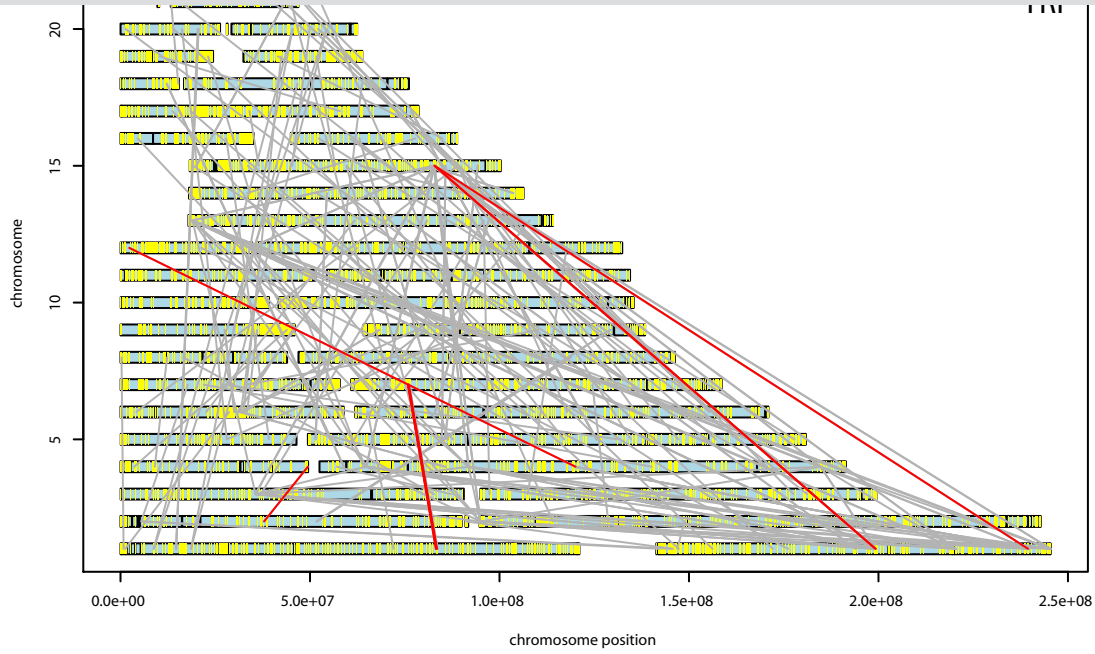
We have shown (see main text) that recombination rates differ significantly between gene ontology classes. Because we have previously identified short DNA sequence motifs that strongly influence recombination activity, we can ask whether the differences in estimated recombination rate reflect differences in the motif density between gene ontology classes. Using the same categories of gene ontology as analysed in the main text we find a strong positive correlation between estimated recombination rate and motif density, suggesting that differences in the genomic density of hotspot-associated motifs are the primary determinant of differences in recombination rate among genes of different molecular function.

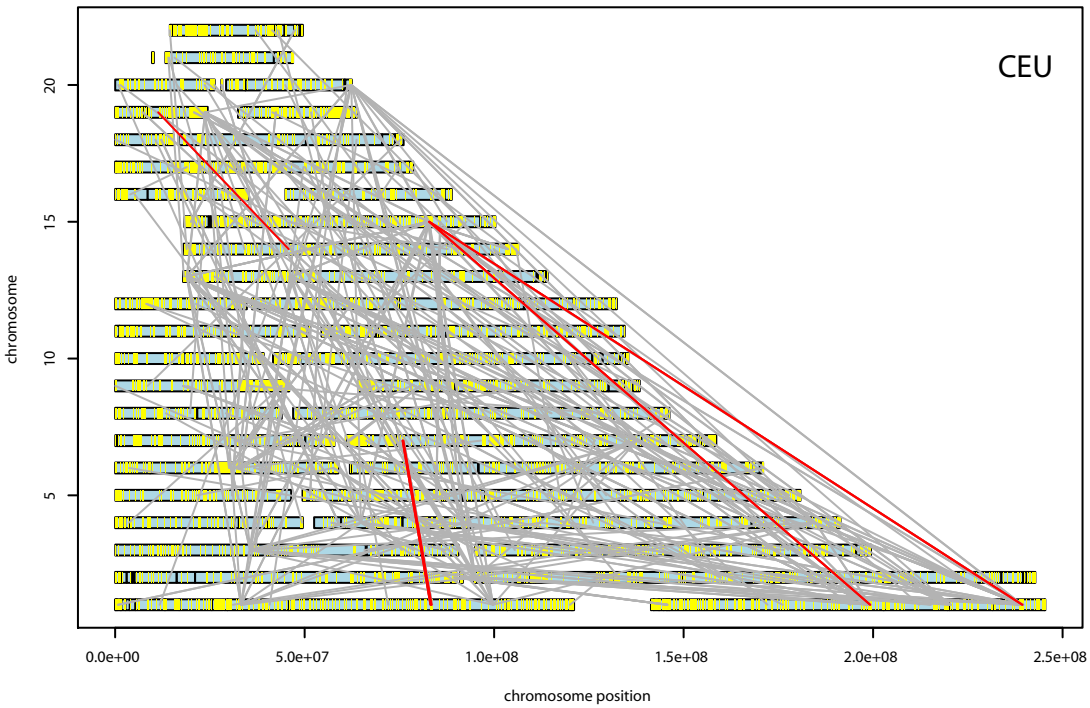


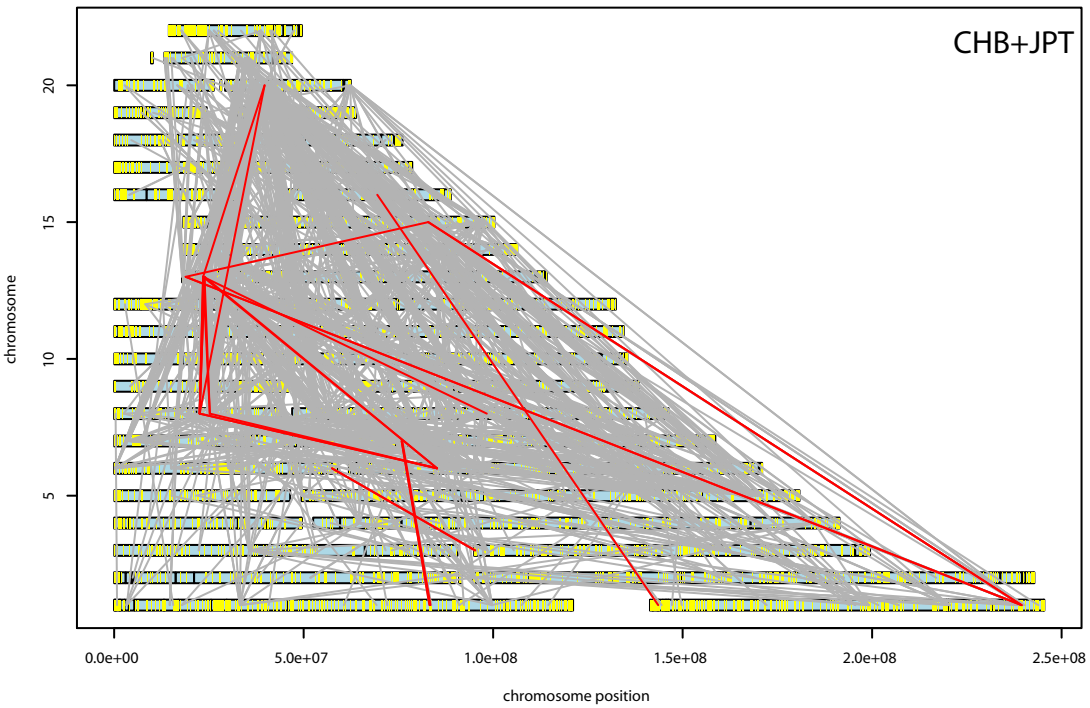
Supplementary Figure 1



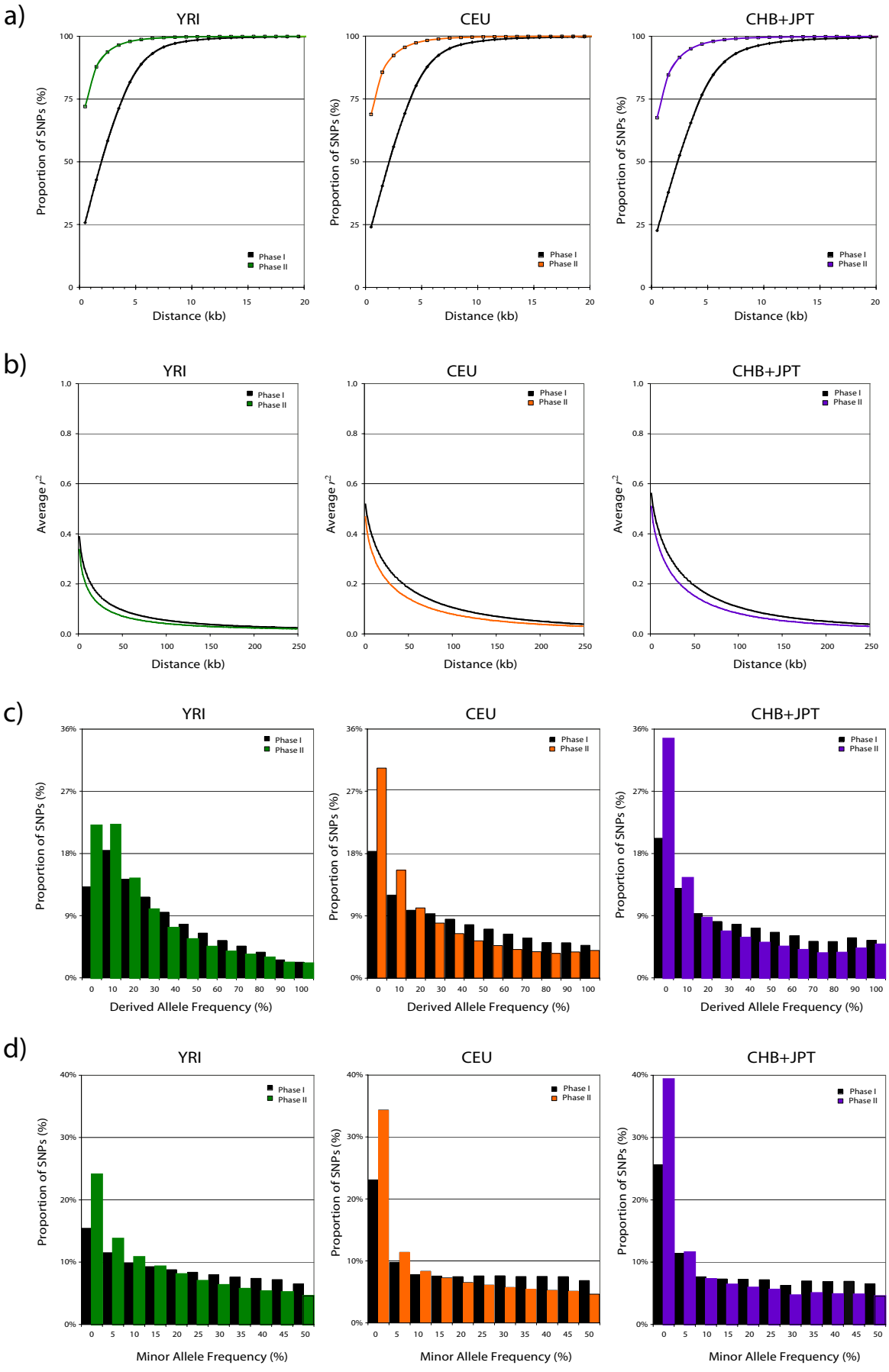
Supplementary figure 2



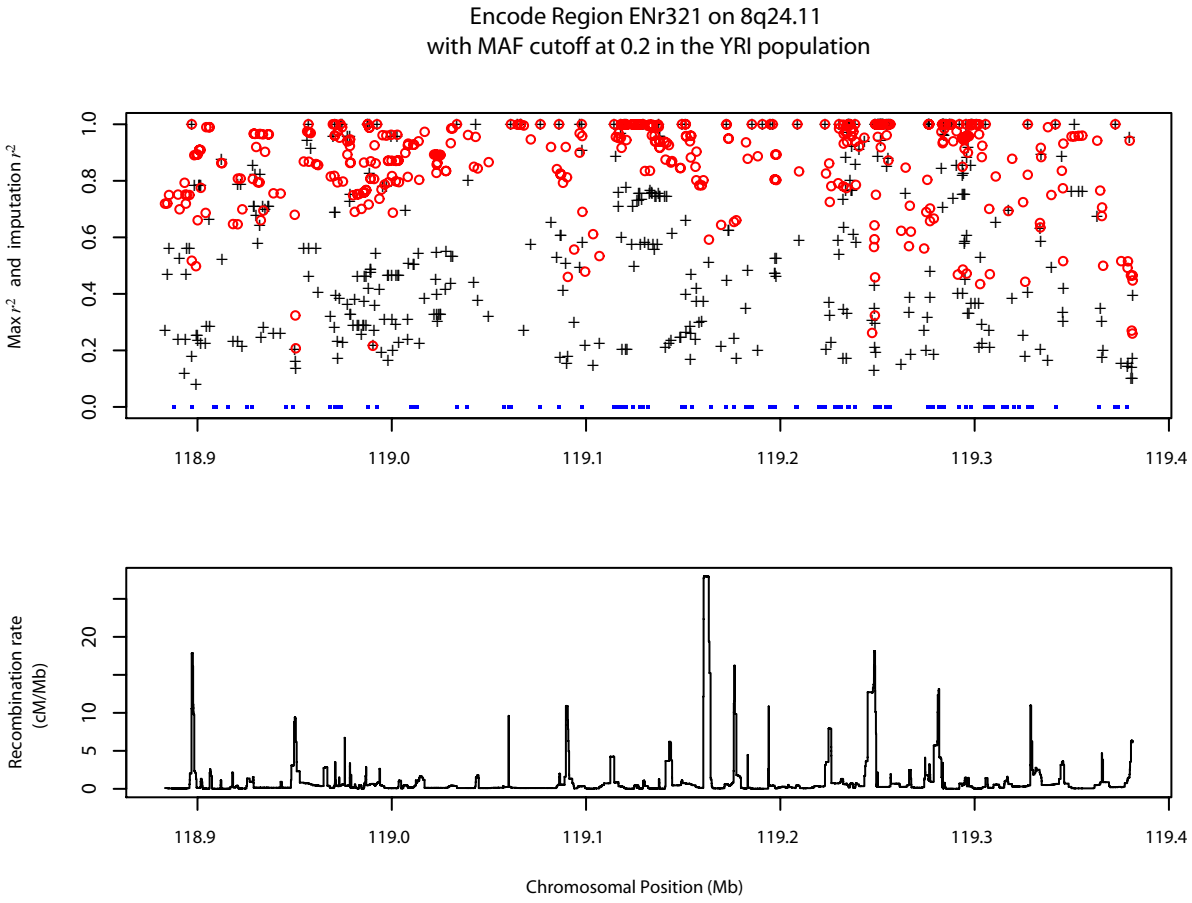




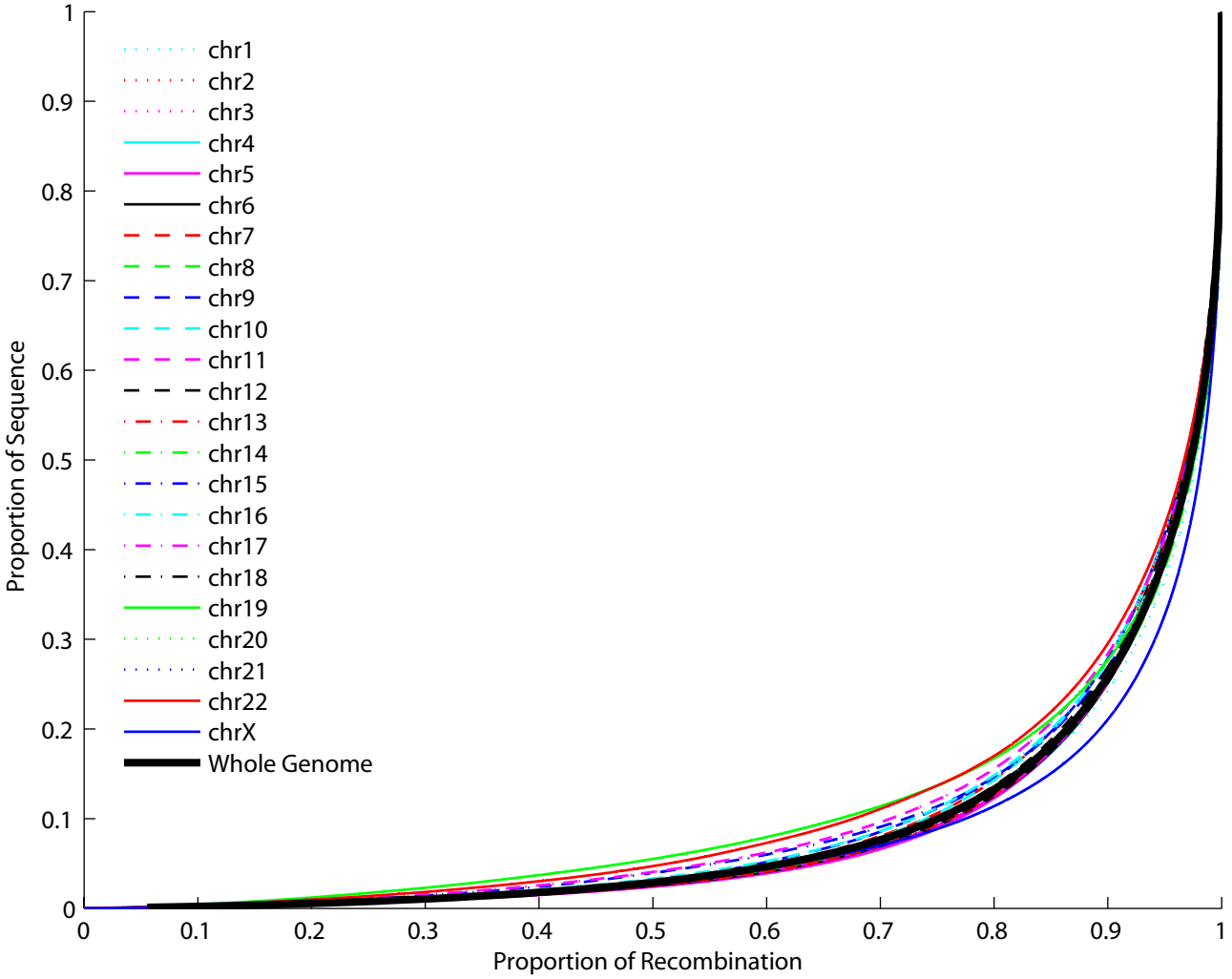
Supplementary Figure 3



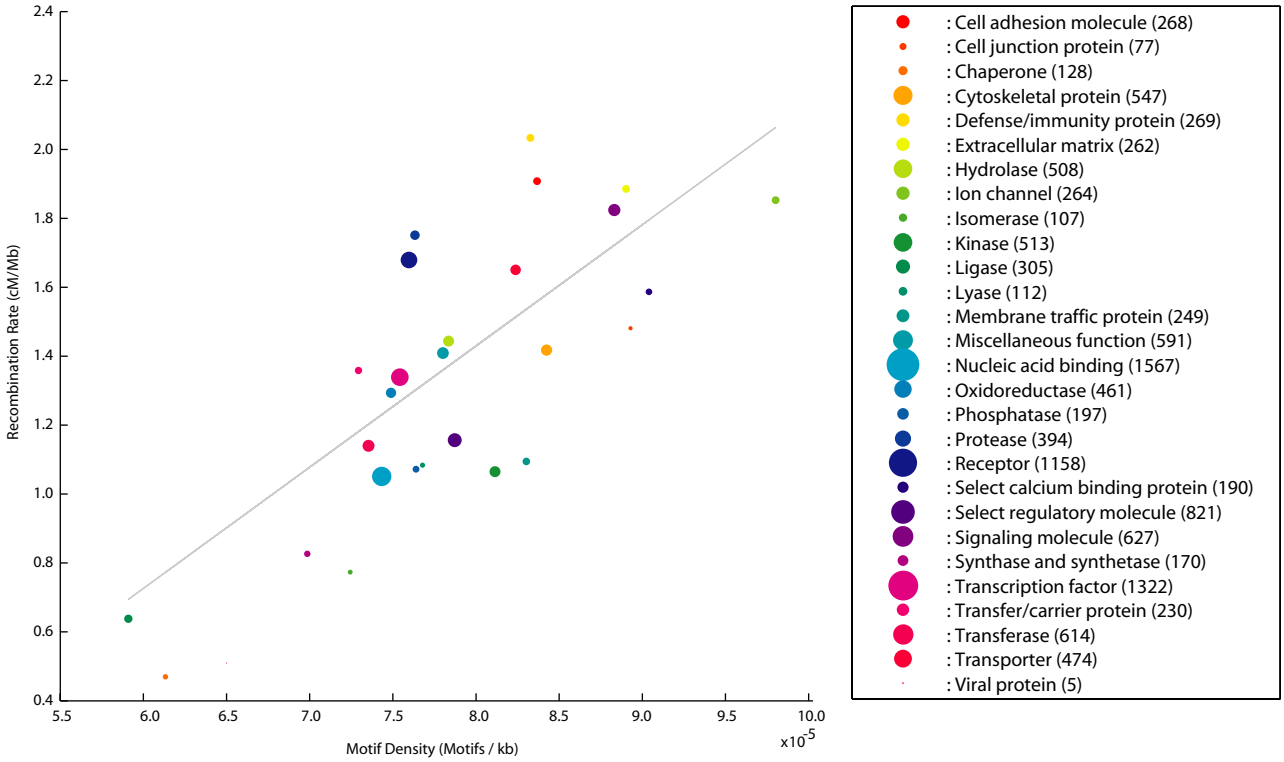
Supplementary Figure 4



Supplementary FIgure 5



Supplementary Figure 6



Supplementary Figure 7