SUPPLEMENTARY MATERIAL

Singapore Genome Variation Project: A haplotype map of three South-East Asian populations

CONTENTS

1 Data preparation

1.1 Sample collection

Subjects enrolled in the Singapore Genome Variation Project (SGVP) were originally recruited for an inter-population study on the genetic variability to drug response, where 100 individuals from each of the Chinese, Malay and Indian population groups were anonymously and randomly chosen from the manifest to partake in SGVP, with only gender and population membership. Of these 300 samples, genomic DNA samples for 99 Chinese, 98 Malay and 95 Indians were chosen for genotyping. Population membership was ascertained on the basis that all four grandparents belong to the same population. Ethical consent for the original study on drug response and further ethical approval for the extension to genome-wide genotyping were granted by two independent Institutional Review Boards at the National University Hospital (Singapore) and the National University of Singapore respectively.

1.2 Genotype data

Genomic DNA for 293 individuals was assayed on the Affymetrix SNP6.0 Genotyping Chip and Illumina1M-single DNA Analysis BeadChip. Three subjects (one from each population) were deliberately genotyped twice for QC purposes, and one was a control individual which was removed from the data after genotype calling. The Affymetrix array yielded data for 934,968 genetic variants including 3,022 control probes, while the Illumina array yielded data for 1,072,820 genetic variants. There were 7 repeats done on the Affymetrix array due to failure to exceed the DM call rate of 86% on the 3,022 control probes, of which one was eventually discarded as the repeated genotyping failed again to make the 86% cut-off.

1.3 Genotype Calling

For Affymetrix, CEL files for 295 samples were submitted for calling (291 individuals, 3 repeats, 1 positive control). Genotypes were called by the BirdSeed calling algorithm [1] from Broad and available in Affymetrix Power Tools apt-1.8.6 (released March 4, 2008). Model files were based on version 2.6 and na24 of the Product files. For Illumina, genotypes for 296 samples were assigned by the proprietary calling algorithm *GenCall* [2, 3] in the BeadStudio Suite by Illumina using the clusterfiles provided by Illumina. A threshold of 0.15 was implemented on the GC score to decide on the confidence of the assigned genotypes: any genotype with a GC score ≥ 0.15 will be accepted while a genotype with a GC score < 0.15 will be rejected and a NULL genotype assigned instead.

1.4 Preliminary SNP QC

A preliminary round of QC was performed on the SNPs from the autosomal chromosomes to identify a set of 'pseudo-cleaned' SNPs for sample QC. This was performed independently for the Affymetrix and Illumina datasets. Five criteria in the stated order were used to identify Affymetrix SNPs for exclusion: (i) missingness > 5% (55,993 SNPs); (ii) HWE significance across all 294 samples $< 10^{-8}$ (2,600 SNPs); (iii) monomorphic SNPs across all 294 samples (67,602 SNPs); (iv) more than 1 discordant genotype across the 3 pairs of duplicated samples (88 SNPs); (v) problems in annotations (36 probes: 28 probes without flanks, 2 pairs of probes mapping to the same rsID but with different flanks (SNP_A-8387337, SNP_A-8388040, SNP_A-8493668, SNP_A-8497683), 2 probes not annotated correctly (SNP_A-1864388, SNP_A-4251461) and 2 probes mapped to the same position but different flanks (SNP_A-2144818, SNP_A-8548122)). This removed a total of 126,309 SNPs out of the total of 892,577 autosomal SNPs. For Illumina, only the first four criteria were used, removing: (i) 33,775 SNPs due to missingness; (ii) 2,475 SNPs due to gross departure from HWE; (iii) 121,327 monomorphic SNPs; (iv) 5 SNPs with discordant genotypes between duplicated samples. In total, 157,582 SNPs out of the total of 1,029,591 autosomal SNPs were removed.

1.5 Sample QC

The quality of the genotype data for each sample was assessed using the SNPs that remained after the preliminary round of SNP QC. This was performed independently for the Affymetrix and Illumina datasets. Samples were identified for removal on the basis of: (i) missingness > 2% (2 samples for Affymetrix, 5 samples for Illumina); (ii) excessive identity-by-state (IBS) genotypes (9 samples for Affymetrix, 10 samples for Illumina – see tables below) where in each identified relationship the sample with the lower missingness is retained. **Tables S9** and **S10** show the extent of IBS between samples on the Affymetrix and Illumina arrays respectively.

The Singapore Genome Variation Project is founded on the basis that there exist genetic differences between subjects from the three populations. As such, there is a need to investigate the genetic evidence of this basis and this is achieved through the use of principal components analysis using the program *pca* distributed with *eigenstrat* [1] (see Section 2 on population structure). SGVP aims to describe the genetic variation found between the three populations, and subjects were recruited into the study to minimize intra-population genetic heterogeneity by confirmation that the parents and both sets of grandparents belonged to the same population. Samples that displayed either evidence of admixture, or clear evidence of discordance between self-reported and genetically inferred population membership are identified and excluded from the study. This is visually assessed from the plots of the informative principal components, which identified seven samples for exclusion. Both Affymetrix and Illumina data identified the same seven samples and **Figure S8** shows the PCA plots for the Illumina data where the seven excluded samples have been circled (see **Table S11**).

We also found that the recorded genders for two subjects were discordant with the genetically inferred genders:

- Sample 016_1 was recorded as a male Chinese subject but was genetically inferred as a female Chinese. As this sample was not related nor a duplicate of the remaining samples, this sample was retained in the analysis as a female Chinese.
- Sample 194_1 was recorded as a male Malay subject but was genetically inferred as a female Chinese. This sample was removed due to a misspecification between the reported population and the genetically inferred population.

For Affymetrix, 17 samples were removed out of the possible 294 samples, and the population composition of the remaining 277 samples is: 97 Chinese, 93 Malays, 87 Indians. For Illumina, 21 samples were removed out of the possible 295 samples, and the population composition of the remaining 274 samples is: 97 Chinese, 91 Malays, 86 Indians.

1.6 SNP QC

For each platform, an independent round of SNP QC is performed on all the genetic data, reinstating all the excluded SNPs from the preliminary round of SNP QC. This round of QC is performed on each population group separately, and SNPs are excluded on the basis of: (i) missingness > 5%, which for Illumina also include 23,812 SNPs with only intensity data and no valid genotype data; (ii) $p_{\mathrm{HWE}} < 0.001$; (iii) > 1 discordant genotype across the three pair of duplicated samples. In addition, Affymetrix SNPs were also excluded if there were annotation problems. The number of SNPs excluded for each criterion can be found in **Table S1**.

1.7 SNP strand synchronisation

Illumina and Affymetrix use their own conventions for defining SNP strands. While the use of such conventions serves its purpose in synchronising SNPs across different chips within the Illumina and Affymetrix family, the conventions are not defined for all SNP flanks found in other platforms. A solution to this is to synchronise the SNPs to the forward/plus strand as defined by the NCBI Build 36.1 assembly.

The SNP flanks in both platforms are aligned using the Needle and Wunsch algorithm with their reported positions on the NCBI Build 36.1 assembly and are annotated plus or minus strand. A perfect match (score = 1) occurs when the flanks align perfectly with the reference assembly. A good match is defined as an alignment that scores greater than 0.8 (gap penalty is 0) and has a difference of 0.3 when compared to the score of the reverse complement alignment score. This is to prevent a mis-annotation when the flanks are partially palindromic (in a reverse complement sense). A discrepancy occurs when the alignment is neither a perfect or good match, in such cases, the strand is manually annotated. All alleles are subsequently mapped to the positive/forward (plus) strand.

In addition, we checked the strand annotations provided by Affymetrix. This was not performed on the Illumina SNP annotation file as it is not available. Annotation based on 932,457 perfect and 1655 good matches were concordant except for 2 SNPs (SNP_A-4251461 and SNP_A-1864388). Affymetrix's strand annotations were wrong for these 2 SNPs. We identified 36 SNPs that were discrepant and were manually annotated of which one (SNP_A-1907434) had flanks which clearly do not belong to the position reported. This SNP did not pass QC and thus there was no need to explicitly exclude this SNP.

For SNPs that are present on both platforms, a useful indication of incorrect encoding due to strand flipping is when concordance improves greatly upon flipping the alleles in one dataset. We flipped the SNPs for the 36,025 SNPs with less than perfect concordance and identified SNPs that have an improvement or have at most a declination of 30% over the original concordance. We obtained 12 SNPs based on this definition, with the details shown in **Table S12**.

Of the 12 SNPs, only rs16942821, rs238137, rs348238, rs624307, rs7299820 had a high potential of being incorrectly encoded upon inspection of the genotype calls. rs16942821 was flipped as Illumina and Affymetrix were targeting differing alleles – C/T and A/C. The genotypes for the remaining SNPs in the original genotype files generated from the laboratory are consistent with downstream encoded genotypes; it is probable that the flipping of SNPs originated upstream either in the platforms' software or incorrect assignment of probes on the chip.

We identified 2 SNPs that are probed differently in both platforms (**Table S13**): SNP rs7171243 had 99.23% concordance assuming the alleles T and G are equivalent. SNP rs16942821 was detected as a flipped SNP.

1.8 Genotyping accuracy

An important feature of any public release of genotype data is the quality of genotyping. To evaluate the quality of the released genotypes, we made use of the duplicated samples to provide a cross-validation of the genotyping accuracy. This assessment is made using the SNPs that remain after the second round of SNP QC. For Affymetrix, the concordance was 99.110%, with an overall call rate across 277 samples of 99.65%. For Illumina, the concordance across 3 pairs of duplicated sample was 99.989%. The overall call rate across 274 samples was 99.858%.

1.9 Data merging

The number of samples for each population group that pass QC on both Affymetrix and Illumina platforms is: 96 Chinese; 89 Malays; 83 Indians. All subsequent analyses are generated based on these common samples. The number of post-QC SNPs that are common to both platforms is: 225,017 for Chinese; 224,016 for Malay; 224,293 for Indian. For these SNPs, additional QC was performed to check the concordance of the genotypes for the common samples on both platforms, as well as to check the consistency of the assayed alleles on both platforms. **Table S14** below indicates the number of common SNPs removed for each population which does not meet the threshold when assessing the concordance of the genotypes for the same samples from the Affymetrix and Illumina platforms. A threshold of 95% was subsequently implemented.

Additionally, we removed the 2 SNPs where the mapped alleles (to the +ve strand) for the Affymetrix array were different to the mapped alleles (to the +ve strand) on the Illumina array.

For common SNPs that are not removed, we retained the genotypes from the platform which has a higher call rate, since the extent of missingness is often a good surrogate for genotyping quality. For SNPs with the same extent of missingness (typically when the call rates for both platforms are both 100%), we use the genotypes from the Illumina array. The concordance and call rates of these SNPs are shown in the **Table S15**. (Note that these figures do not include the SNPs with inter-platform concordance < 95%).

The total number of unique autosomal SNPs that pass QC in each of the 3 populations is:

Chinese	: 1,584,040
Malay	: 1,580,905
Indian	: 1,583,454

2 Population Structure

2.1 Samples and genotype data

A total of 2,896,293 SNPs were common to the four HapMap panels in release 26 (as of 17 Dec 08), of which 1,423,464 SNPs were common to all three SGVP populations. We considered every 10th SNP from this set of SNPs common to the seven populations, which yielded 142,347 SNPs for performing population structure analyses. For population structure analysis with the genotype data from the Human Genome Diversity Project (HGDP), 610,437 SNPs were common across all the HapMap, HGDP and SGVP populations. This set was thinned by selecting every 6th SNP, resulting in 101,740 SNPs, for performing principal component analysis (PCA).

2.2 Principal component analysis

Principal component analysis was performed using the *pca* program distributed together with *eigenstrat* [4]. We ran *pca* to produce the first 20 principal components and identified 16 HapMap individuals for removal by the outlier classification criterion. Six sets of analyses were performed:

(i) with the HGDP, HapMap and SGVP populations, consisting of 1,421 individuals;
(ii) with only the seven HapMap and SGVP populations, consisting of 462 individuals;
(iii)with samples of East Asian ancestries, defined as East Asian samples from HGDP, HapMap CHB and JPT, and SGVP CHS, consisting of 409 individuals;
(iv)with the three populations from Far East Asia from the HapMap (CHB, JPT) and SGVP (CHS), consisting of 181 individuals;
(v) within the three SGVP populations, consisting of 268 individuals;
(vi)with the two Chinese cohorts (CHB, CHS), consisting of 138 individuals.

2.3 F_{ST} calculation

We implemented the weighted version of F_{ST} calculation used by the International HapMap Project [5] which accounts for differences in the number of chromosomes in each population. This is given as

$$F_{ST} = 1 - \frac{\sum_j \binom{n_j}{2} \sum_i 2\frac{n_{ij}}{n_{ij}-1} x_{ij}\left(1-x_{ij}\right) \Big/ \sum_j \binom{n_j}{2}}{\sum_i 2\frac{n_i}{n_i-1} x_i(1-x_i)}$$

where:

x_{ij} = the estimated frequency (proportion) of the minor allele at SNP *i* in population *j*;
n_{ij} = the number of genotyped chromosomes at SNP *i*;
n_j = the number of chromosomes analysed in population *j*.

We also calculated the SNP-specific F_{ST} statistic between pairs of populations for every SNP that passes QC using the formula

$$F_{ST} = \frac{\left(p_1 - p_2\right)^2}{\left(p_1 + p_2\right)\left(2 - p_1 - p_2\right)},$$

where p_1 and p_2 denote the frequencies of a specific allele at a SNP in each of the two populations respectively. The pairwise F_{ST} between pairs of HapMap and SGVP populations can be found at **Table S2**.

3 SNP and haplotype analysis

3.1 Comparison of allele frequencies across pairwise panels

Heatmaps of genome-wide allele frequencies are used as visual summary of the allele frequencies distributions across pairs of populations. We consider only the 1,369,502 common and polymorphic SNPs across the three groups, and the minor allele is defined after agglomerating the genotype data from all three populations. For each SNP in a specified population, the frequency of the defined allele is calculated using only the samples from this population. Twenty allele frequency bins each spanning 0.05 units are constructed for each population, and we tabulate the number of SNPs found in each bin. In a comparison of the allele frequency distribution between two populations, we considered 400 allele frequency bins from a 20 × 20 grid. The horizontal axis defines the 20 allele frequency bins for the first population while the vertical axis defines the 20 allele frequency bins for the second population. Each SNP is thus binned according to the allele frequency found in the two populations respectively (see **Fig. S4**).

3.2 Recombination rates estimation

Population specific recombination rates are estimated using the program LDhat (version 2.1). We used the lookup table for 192 chromosomes with $\theta = 0.001$ per site for the Chinese, as well as to generate lookup tables for 178 and 166 chromosomes for the Malay and Indian data respectively. The number of reversible jump MCMC iteration was set at 10,000,000, and a block penalty of 5 was implemented. The thinning interval was set at 2000 such that a sampling is performed every 2000 iterations. The resultant output was summarized using the *stat* program available in LDhat, yielding the mean and median recombination rate and the associated 95% CI for the estimated mean, after excluding the first 100,000 iterations as burn-in. We used the mean recombination rate and the physical distance between consecutive SNPs to calculate the genetic distance for each chromosome for each group.

3.3 Haplotype phasing

The SGVP genotype data was phased using the program *fastPHASE* (version 1.3) [6]. A series of tests was performed to investigate the optimal choice of parameters to be used, given realistic expectations on the running time. We vary the number of haplotype cluster *K* between 6 and 20 inclusive at the default setting of 20 EM runs, and perform independent rounds of phasing with and without incorporating subpopulation labels. Each chromosome was phased independently and we ran 10 iterations of error rate estimations for each chromosome. In each iteration, 1000 consecutive SNPs are randomly selected, of which approximately 10% of the observed genotypes are masked across all individuals considered and imputed by the algorithm. The error rate represents the extent of the discordance between the imputed genotypes and the observed genotypes averaged over 10 iterations. The error rates for the values of *K* considered with and without including subpopulation labeling are shown in **Figure S9**. Based on the empirical error rates, the final phasing was performed separately for each SGVP population with *K* = 14.

4 Analysis of LD and recent positive natural selection

4.1 Linkage disequilibrium analysis

The extent of linkage disequilibrium (LD) between two SNPs is calculated off the phased haplotypes using the program *haploview* [7], and we quantify LD by three metrics: (i) the square of the genetic correlation coefficient r^2; (ii) D'; (iii) the LOD score. Every SNP with minor allele frequency ≥ 5% has a chance to be defined as the focal SNP and for each focal SNP, we compute the LD between the focal SNP and all other SNPs with MAF ≥ 5% that are found within 250kb upstream and downstream of the focal SNP.

4.2 Analysis of LD variation

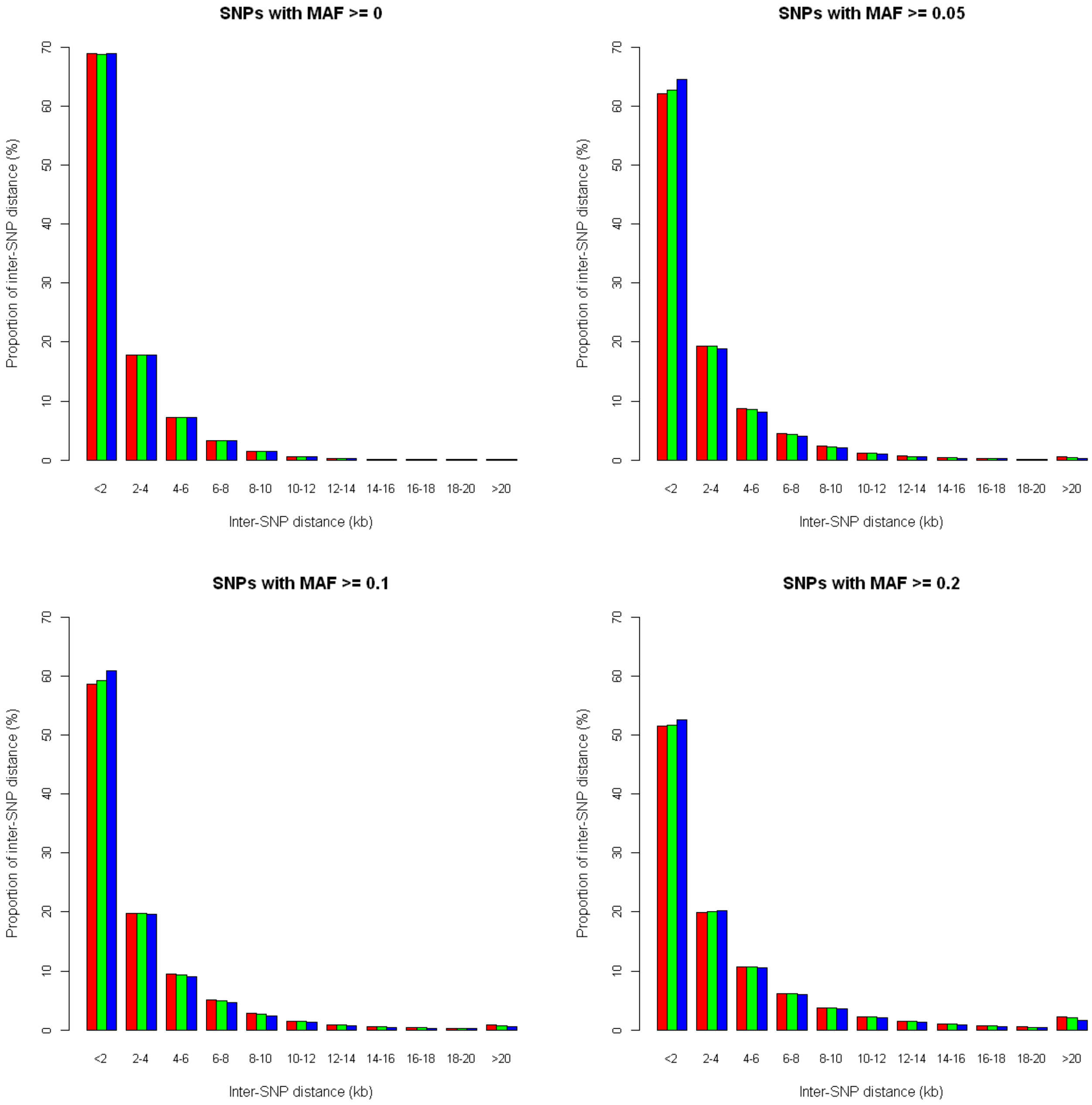
Comparison of regional LD between two populations was performed with the *varLD* algorithm [8]. Briefly, we consider windows of 50 consecutive SNPs common to both populations, and calculate the signed r^2, defined as the r^2 with the sign of the D' metric, between all possible pairs of these SNPs. Consequently, we construct a 50 × 50 symmetric matrix for each population where the $(i, j)^{th}$ element represents the signed r^2 metric between the i^{th} and j^{th} SNPs calculated. We compare the equality between the two matrices by comparing the extent of departures between the eigenvalues of these matrices. This is given by the sum of the absolute difference between the ranked eigenvalues for the two matrices, and this constitutes a score for each window of 50 SNPs. The extent of LD differences in each window is assessed by comparing the relative rank of the score obtained against the distribution of scores in the genome, and we identify regions which constitute the top 5% of the distribution of the scores. For visualizing the signals from comparisons across multiple population-pairs, we standardized the scores to have a mean of zero and a standard deviation of one. To avoid excessively long tables, we show only the top 0.1% of the distribution for comparisons between all possible pairs of SGVP populations, CEU with each SGVP population, and between CHS and both CHB and JPT+CHB in **Table S4**.

4.3 Detecting signatures of positive selection

In setting up the analysis with the integrated haplotype score (iHS), we first define the extended haplotype homozygosity (EHH) as the probability of identity-by-descent for two randomly chosen haplotypes that are carrying the core haplotype of interest within an interval around the core region [9]. The EHH is calculated for each SNP, and the iHS is calculated up to an EHH score of 0.05 unless we encounter a gap between adjacent SNPs of greater than 200kb. For adjacent SNPs with gaps of between 20kb and 200kb, the scaling factor described by Voight and colleagues [10] was implemented to correct for the artificial inflation of the calculated iHS. The iHS was not calculated for SNPs if at least one of the following conditions were encountered: (i) minor allele frequencies < 5%; (ii) the derived allele was unknown or did not agree with either of the two possible alleles defined for the SGVP data; (iii) if the EHH did not drop below 0.05 within 2.5Mb. The designations for the derived alleles were obtained from the Haplotter website, and the recombination rates that were averaged over all the HapMap populations were used. The obtained iHS statistics were normalized in 20 derived allele frequency bins, each spanning 5%. Candidate regions of positive selection are identified by a clustering of SNPs with high iHS values, defined as |iHS| > 2. Regions of selection can be identified by SNPs with high iHS scores or by identifying regions of the genome with an unusual density of high iHS scores. To identify the latter we calculated the proportion of SNPs with iHS > 2.0 in all 100 kb non-overlapping windows and identified windows with the top 1% proportion of significant SNPs. Windows with a total of less than ten SNPs were dropped from the regional analysis.

The XP-EHH test compares the evidence of selection across two populations at a core SNP for a stated direction in each chromosome. Briefly, given a core SNP and the direction of analysis, only SNPs found in both populations and within 1Mb of the core SNP are considered. A test is only valid if there is at least one SNP in this region with an EHH of between 0.03 and 0.05, calculated with respect to all chromosomes in both populations. When there is more than one SNP that satisfies this criterion, the SNP with an EHH closest to 0.04 is considered. At each population, the integral of the EHH at all SNPs between the core SNP and this latter SNP is taken. The XP-EHH log-ratio is defined by the logarithm of the ratios of the integrals from both populations. The collection of XP-EHH log-ratios for every pair of populations is standardized such that the resultant distribution has zero mean and unit variance (**Figure S10**). A clustering of extreme positive values of these standardized scores suggests that a selection event is likely to have occurred in one population but not the other, whereas extreme negative values suggest a selection event in the latter population but not the former. We primarily use XP-EHH to confirm differential iHS signals across different populations, and thus we implement a comparatively liberal threshold, defining a candidate selection region as one with a cluster of SNPs with absolute XP-EHH scores > 2.5.

5 Supplementary figures


SNPs with MAF >= 0

SNPs with MAF >= 0.05

SNPs with MAF >= 0.1

SNPs with MAF >= 0.2

Figure S1. Distribution of inter-SNP distance
The inter-SNP distance for each population (red for CHS, blue for INS and green for MAS), for SNPs binned by minor allele frequencies ≥ 0, ≥ 0.05, ≥ 0.10 and ≥ 0.20.

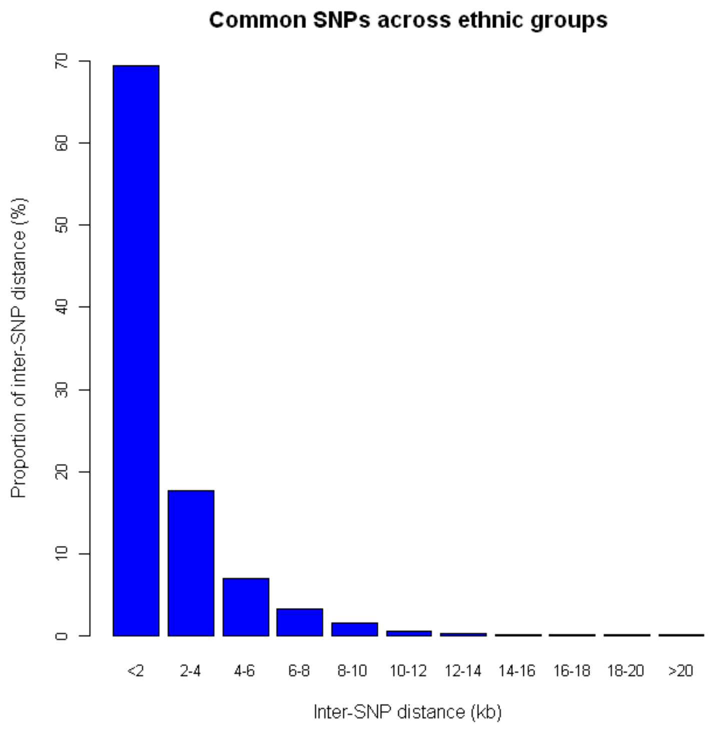


Figure S2. The distribution of the inter-SNP distance for the 1,369,502 post-QC SNPs that are polymorphic and common across all three SGVP populations

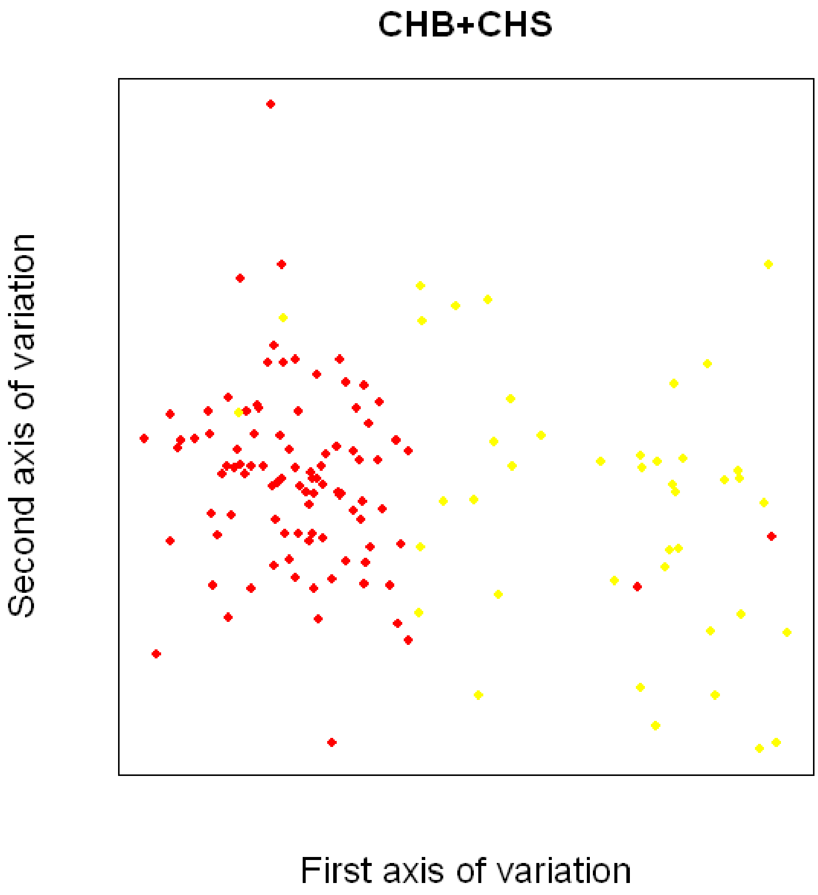


Figure S3. PCA plots for the two Chinese populations

The figure plots the first two axes of variation when the PCA only considers samples from Singapore Chinese (red) and HapMap Han Chinese in Beijing, China (yellow).

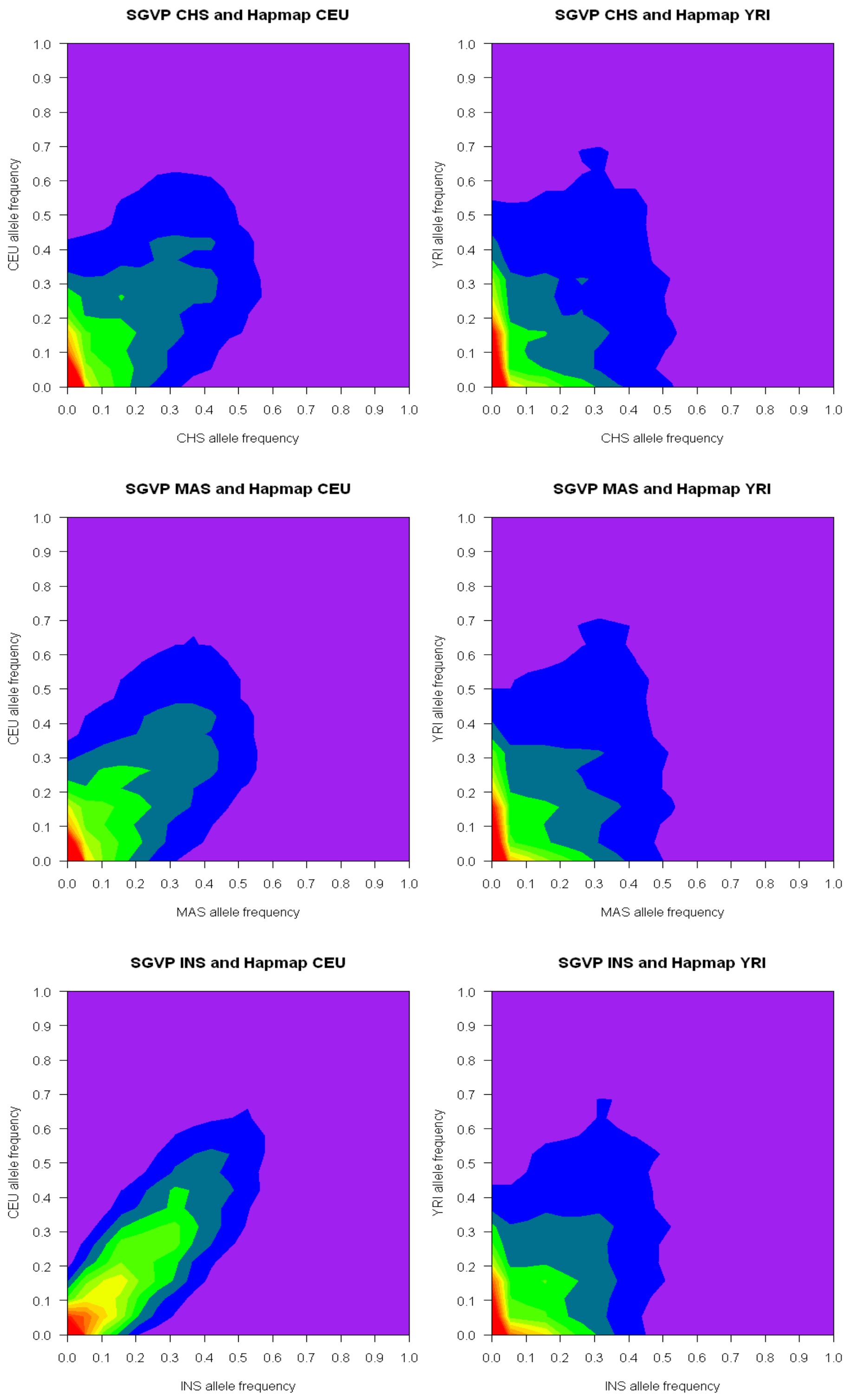


Figure S4. Allele frequency comparison between SGVP populations and CEU (left column) and YRI (right column)

The axes in each figure represent the allele frequencies in two populations. In each pair of comparison, we calculate the frequency of the minor allele for each SNP after agglomerating the genotype data from all three populations. The intensity of the colour represents the number of SNPs that display the corresponding allele frequencies in the two populations, in 20 bins each of width 0.05. The colour legend follows that of **Figure 2**. For example, purple regions indicate that very few SNPs are observed to possess the allele frequency combination corresponding to values represented by the appropriate x- and y-axis markings.

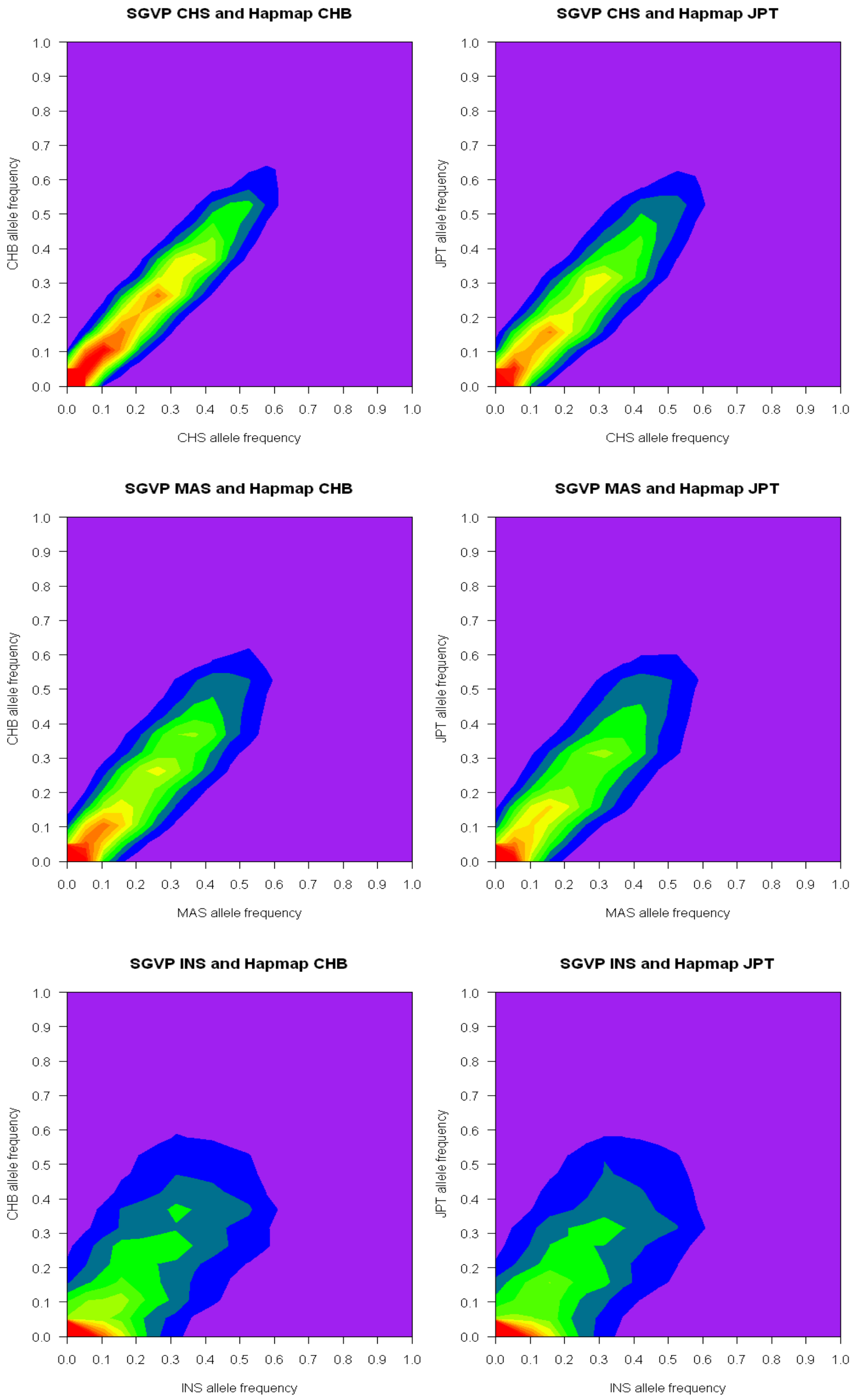


Figure S4 (continued). Allele frequency comparison between SGVP populations and CHB (left column) and JPT (right column)

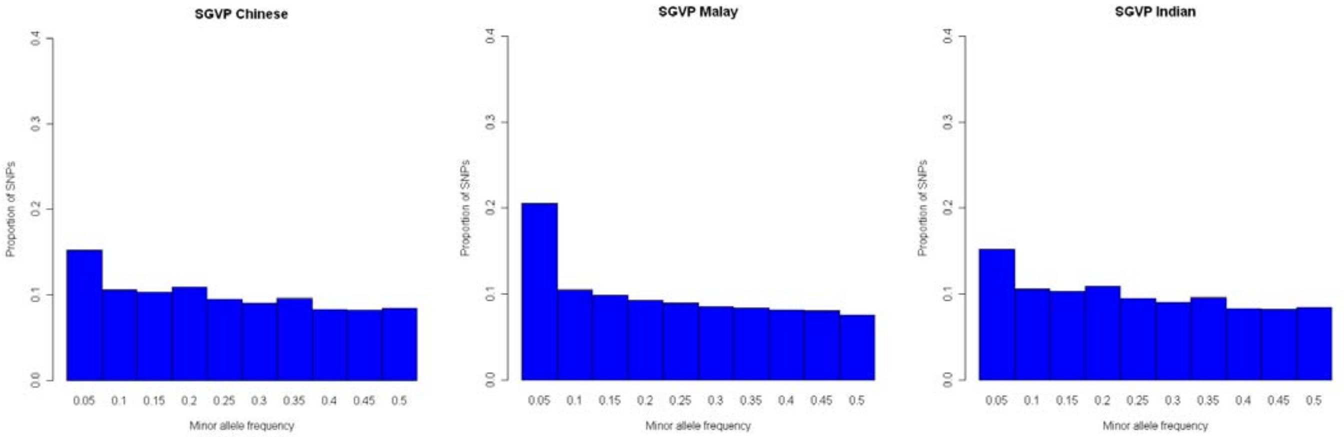


Figure S5. Distribution of minor allele frequencies
Histograms for the allelic spectrum for each SGVP population across the assayed SNPs by placing the SNPs into minor allele frequency bins of width 0.05.

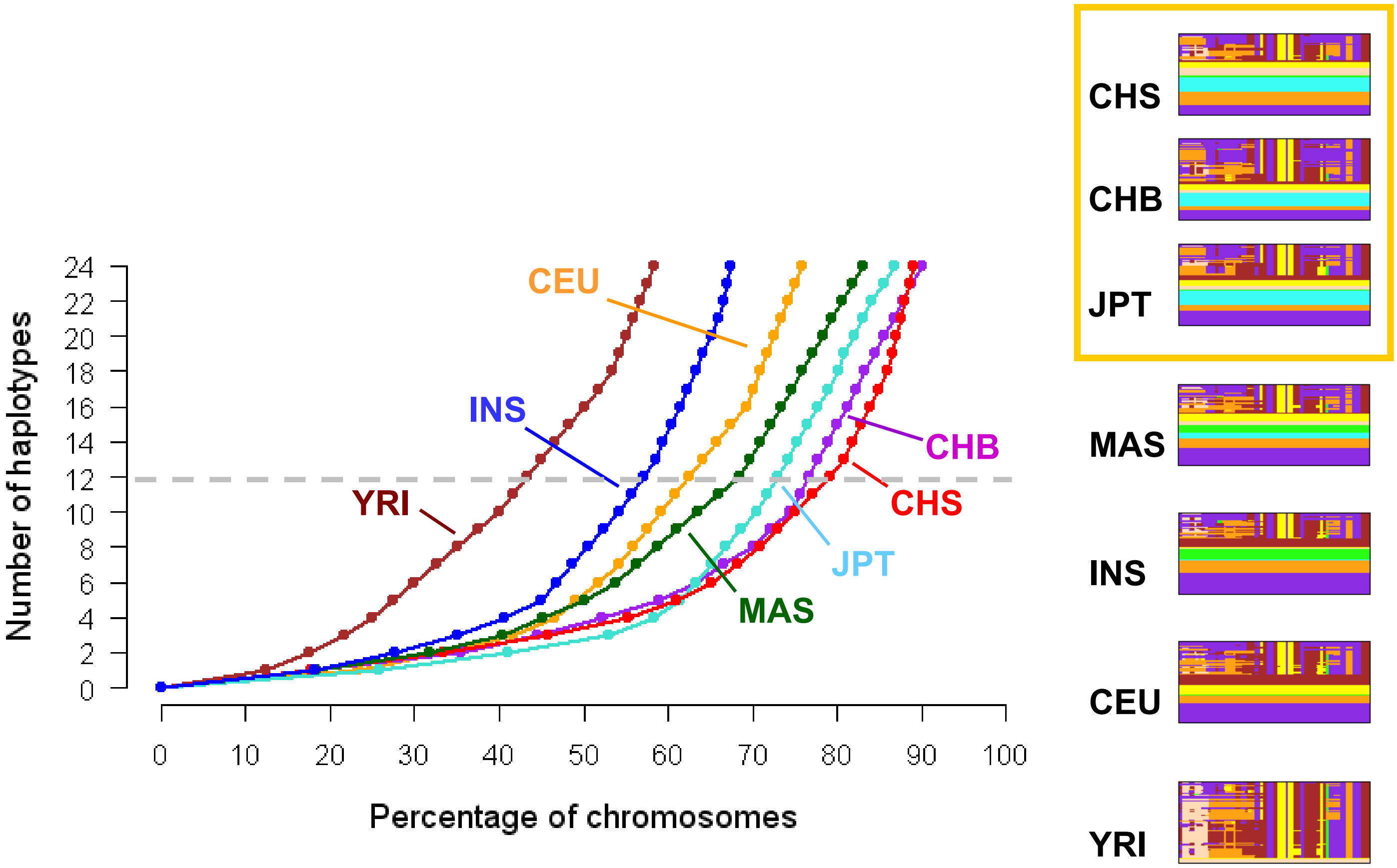
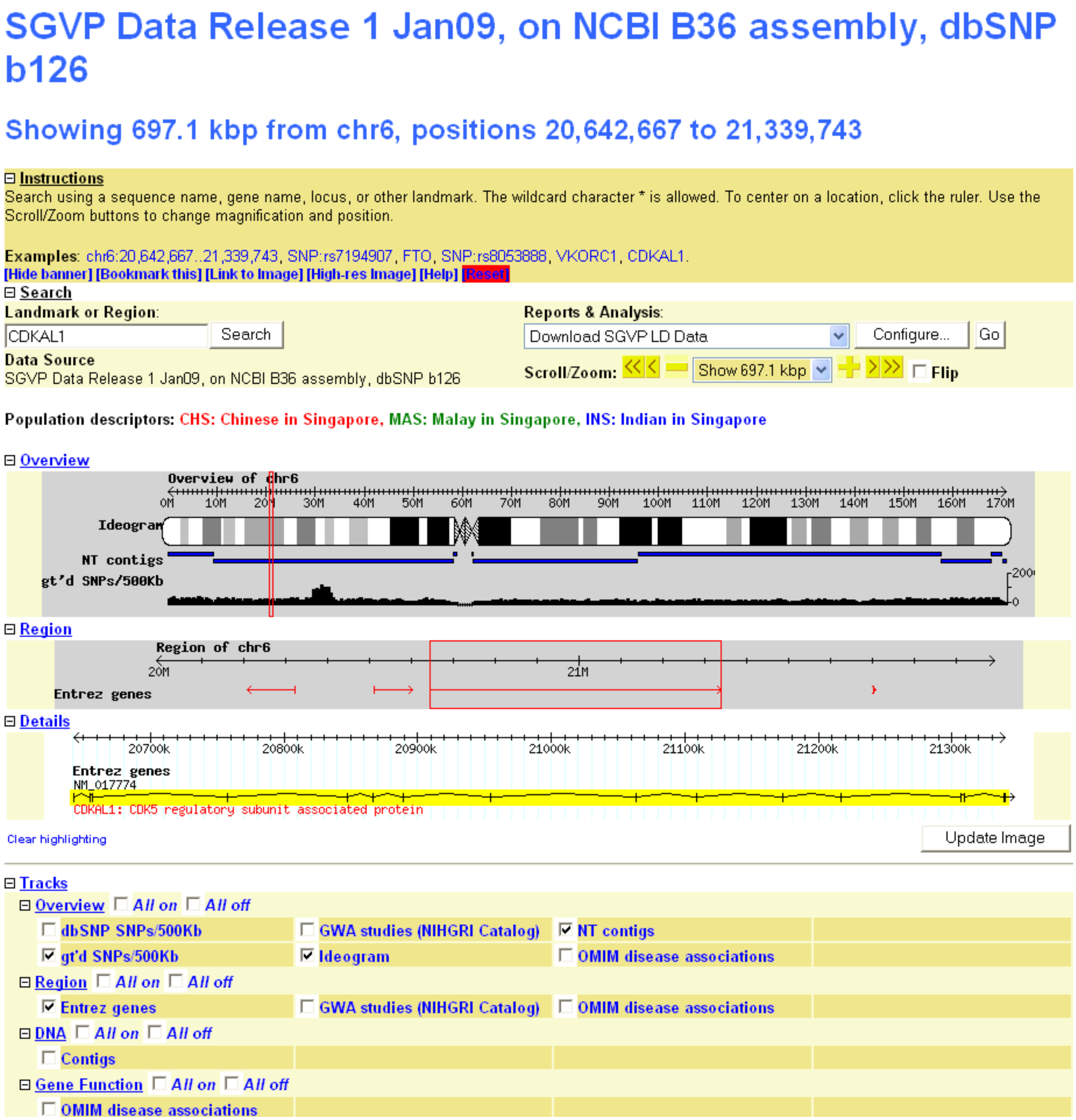


Figure S6. Haplotype diversity across the seven SGVP and HapMap populations
The graph shows the average percentage of the chromosomes within each population that can be accounted for by the corresponding number of distinct haplotypes. This analysis considers 22 unlinked regions of 500kb from each of the autosomal chromosomes, spanning an average of 174 SNPs. The horizontal grey dashed line is drawn at 12 haplotypes for interpretation of haplotype diversity in the main text. The seven panels on the right indicate the extent of haplotype sharing across the 500kb region on chromosome 1, where seven canonical haplotype forms are identified and the chromosomes from each population is mapped either uniquely to, or as a mosaic of, these seven canonical haplotypes. The canonical haplotypes correspond to the haplotype forms which most of the chromosomes are similar to, and each canonical haplotype is assigned a unique colour scheme. The three populations with ancestries from the Far East (CHB, CHS and JPT) have been boxed, and it is evident that the haplotype diversity is considerably similar across these three populations relative to the rest of the populations.



SGVP Data Release 1 Jan09, on NCBI B36 assembly, dbSNP b126

Showing 697.1 kbp from chr6, positions 20,642,667 to 21,339,743

⊟ Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: chr6:20,642,667..21,339,743, SNP:rs7194907, FTO, SNP:rs8053888, VKORC1, CDKAL1.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help] [Reset]

⊟ Search

Landmark or Region: CDKAL1 [Search]

Data Source

SGVP Data Release 1 Jan09, on NCBI B36 assembly, dbSNP b126

Reports & Analysis: Download SGVP LD Data [Configure...] [Go]

Scroll/Zoom: << < − Show 697.1 kbp + > >> □ Flip

Population descriptors: CHS: Chinese in Singapore, MAS: Malay in Singapore, INS: Indian in Singapore

⊟ Overview

⊟ Region

⊟ Details

Clear highlighting

[Update Image]

⊟ Tracks

⊟ Overview □ All on □ All off

□ dbSNP SNPs/500Kb	□ GWA studies (NIHGRI Catalog)	☑ NT contigs	
☑ gt'd SNPs/500Kb	☑ Ideogram	□ OMIM disease associations	

⊟ Region □ All on □ All off

☑ Entrez genes	□ GWA studies (NIHGRI Catalog)	□ OMIM disease associations	

⊟ DNA □ All on □ All off

□ Contigs			

⊟ Gene Function □ All on □ All off

□ OMIM disease associations			

Figure S7. Screen capture of the SGVP genome browser

A screen capture of the publicly available genome browser at http://www.nus-cme.org.sg/SGVP/ for accessing, viewing and downloading data from the SGVP resource.

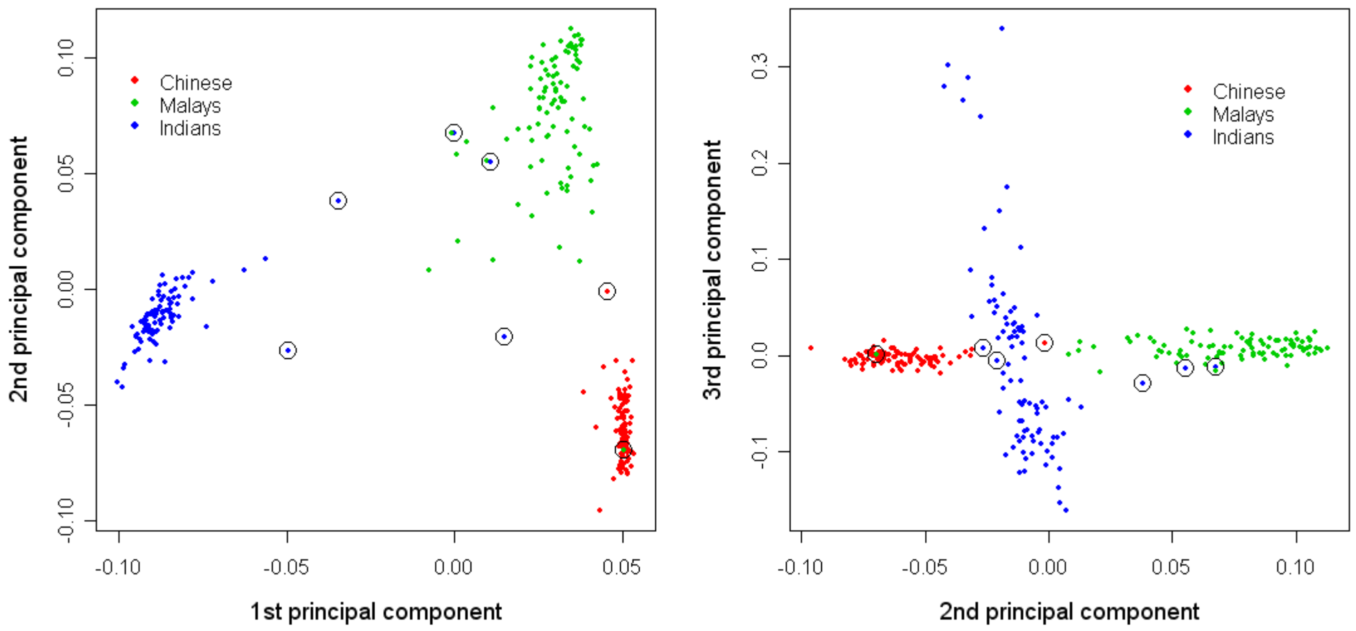


Figure S8. PCA plots to identify samples with discordant self-reported and genetically inferred population membership

Plots of the first two principal components (left) and the second-third principal components (right) for identifying samples from the three SGVP populations where the self-reported and genetically inferred population memberships are discordant. Seven samples have been visually identified for exclusion and are circled in the plots above.

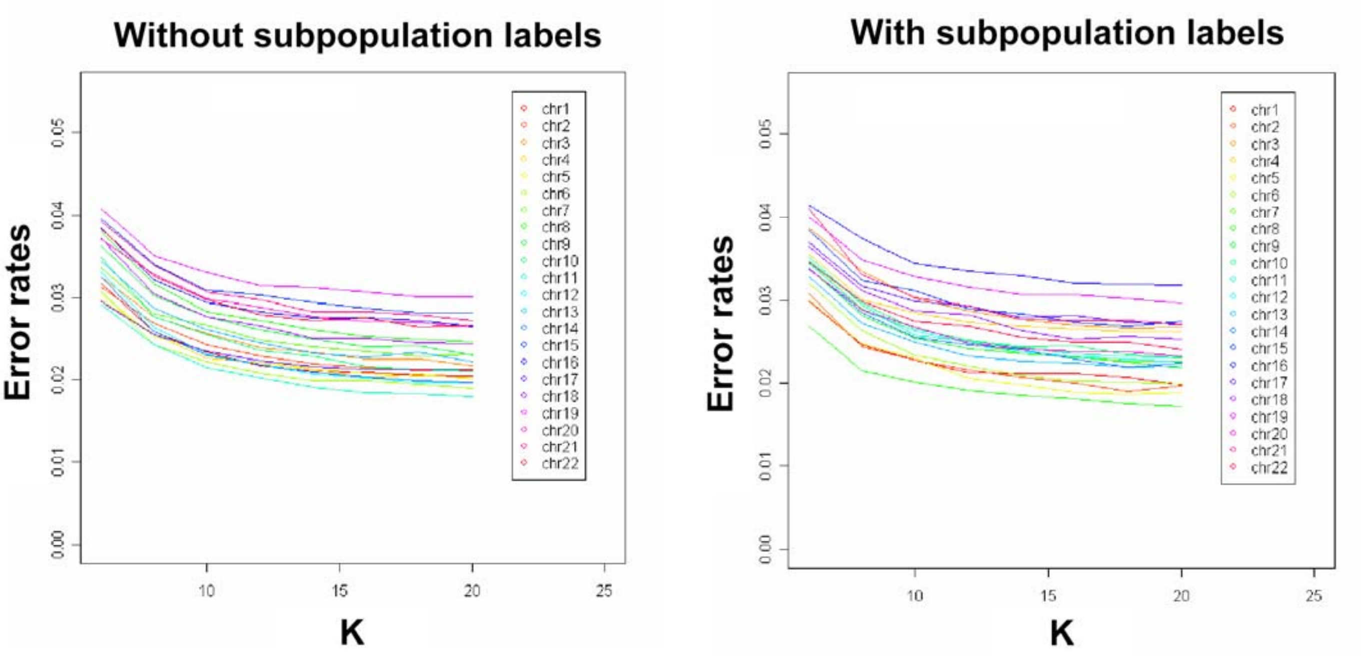


Figure S9. Haplotype phasing error rates at different selection of the number of haplotype cluster *K*

Assessment of the performance of haplotype phasing at each chromosome assuming different number of haplotype clusters. The left plot shows the performance without incorporating population labels across the SGVP populations, while the right plot uses information on the population membership of each chromosome during haplotype phasing.

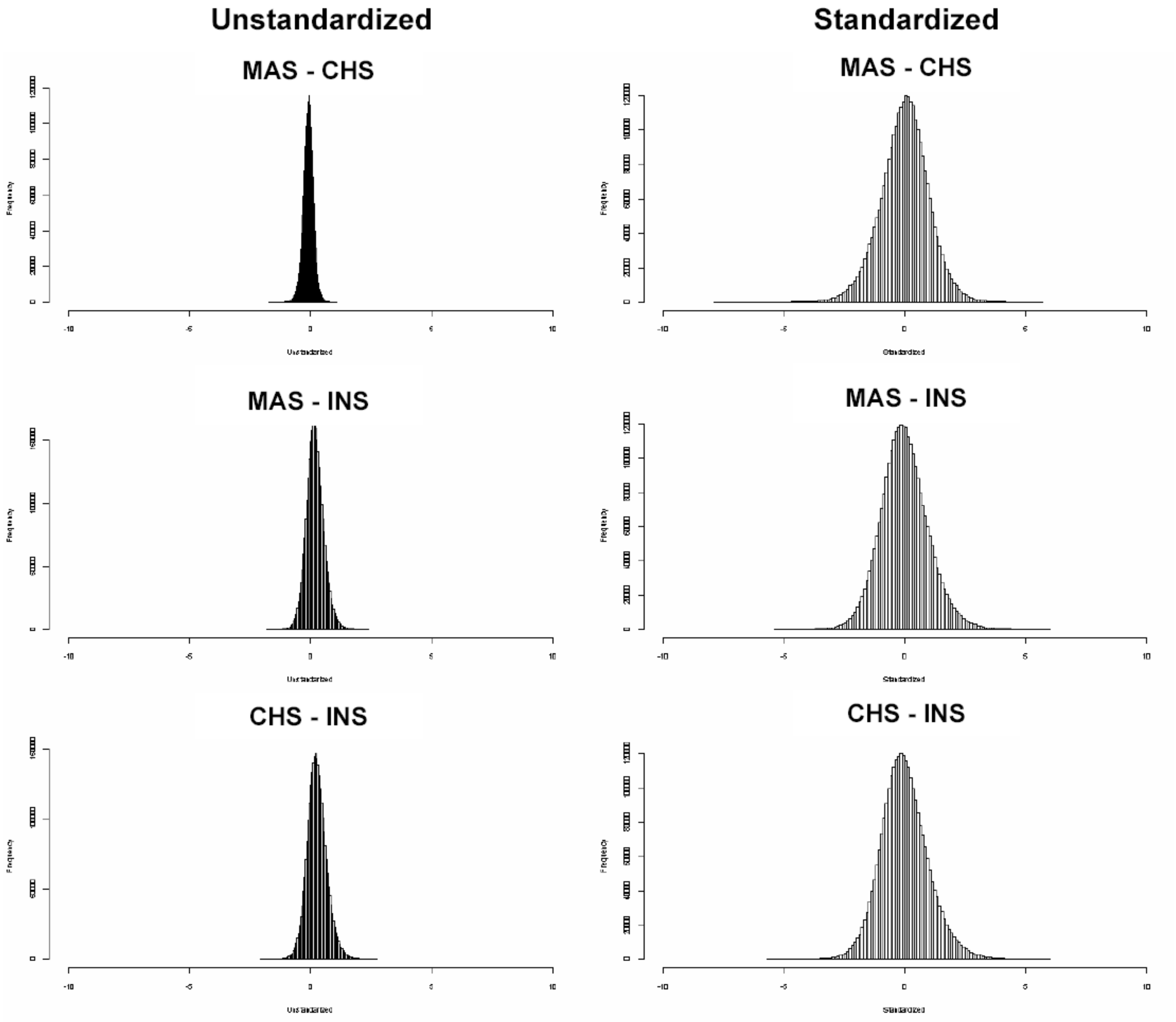


Figure S10. XP-EHH values before and after standardization
To allow for comparison between different population pairs, the collection of XP-EHH log-ratios for every pair of populations is standardized to have zero mean and unit variance.

6 Supplementary tables

Table S1. Data quality control outcome after excluding the three clinical duplicates

Criteria	**Affymetrix**			**Illumina**		
	CHS	**MAS**	**INS**	**CHS**	**MAS**	**INS**
Sample QC						
> 2% missing data	1	0	1	1	2	2
Excessive IBS	0	4	2	0	4	2
Discordant membership	1	1	5	1	1	5
Samples remaining	97	93	87	97	91	86
	CHS	**MAS**	**INS**			
Samples remaining on both arrays	96	89	83			
Autosomal SNP QC						
> 5% missing data	56,683	59,482	57,429	34,143*	33,889*	32,037*
HWE *P*-value < 0.001	3,261	4,060	4,550	3,120	3,886	4,365
> 1 discordant genotype	86	84	93	5	6	6
Annotation failures	34	34	34	0	0	0
SNPs remaining	832,513	828,917	830,471	992,323	991,810	993,183
	CHS	**MAS**	**INS**			
SNP merging QC						
< 95% concordance	482	552	536			
Discordant alleles	2	2	2			
Autosomal SNPs remaining on both arrays	1,584,040	1,580,905	1,583,454			

* Inclusive of 20,400 intensity-only probes that gave NULL calls.

Table S2. F_{ST} calculation for pairs of populations between the seven HapMap and SGVP populations

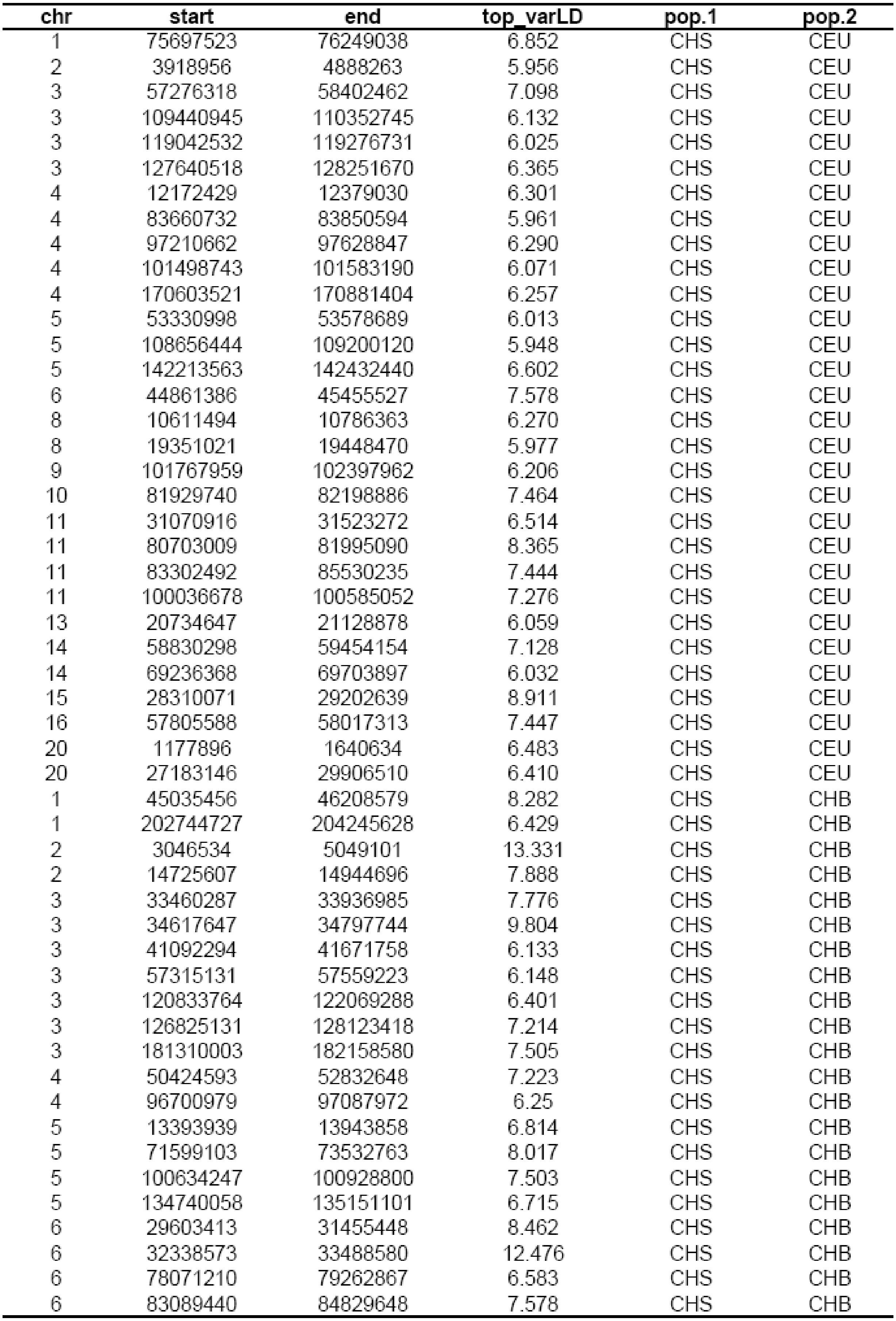
Population A	Population B	HAPMAP Fst	SNP-specific Fst				
			Mean	95% C.I.	1st Quartile	Median	3rd Quartile
CEU	CHB	0.0502	0.0534	(0.0533, 0.0535)	0.0070	0.0289	0.0738
CEU	JPT	0.0505	0.0545	(0.0544, 0.0546)	0.0072	0.0300	0.0757
CEU	CHS	0.0629	0.0532	(0.0532, 0.0534)	0.0069	0.0283	0.0740
CEU	INS	0.0164	0.0199	(0.0198, 0.0199)	0.0024	0.0095	0.0268
CEU	MAS	0.0508	0.0450	(0.0449, 0.0451)	0.0056	0.0235	0.0624
CEU	YRI	0.0770	0.0711	(0.0710, 0.0712)	0.0119	0.0432	0.1010
CHB	CHS	0.0015	0.0050	(0.0049, 0.0050)	0.0006	0.0025	0.0064
CHB	JPT	0.0034	0.0090	(0.0090, 0.0090)	0.0010	0.0048	0.0118
CHB	INS	0.0206	0.0351	(0.0350, 0.0352)	0.0045	0.0178	0.0483
CHB	MAS	0.0031	0.0108	(0.0107, 0.0108)	0.0013	0.0054	0.0142
CHB	YRI	0.0852	0.0817	(0.0815, 0.0818)	0.0143	0.0495	0.1158
JPT	CHS	0.0043	0.0086	(0.0086, 0.0086)	0.0010	0.0041	0.0114
JPT	INS	0.0203	0.0361	(0.0360, 0.0361)	0.0048	0.0184	0.0500
JPT	MAS	0.0053	0.0137	(0.0136, 0.0137)	0.0018	0.0063	0.0179
JPT	YRI	0.0851	0.0827	(0.0826, 0.0829)	0.0144	0.0495	0.1178
MAS	YRI	0.0871	0.0747	(0.0745, 0.0748)	0.0128	0.0443	0.1070
INS	MAS	0.0274	0.0260	(0.0259, 0.0261)	0.0030	0.0129	0.0357
INS	YRI	0.0663	0.0633	(0.0632, 0.0635)	0.0097	0.0378	0.0901

Calculations are based on 1,423,464 SNPs that are common to all seven populations.

Table S3. Top 10 candidate regions of LD variation in each of SGVP populations against HapMap CEU

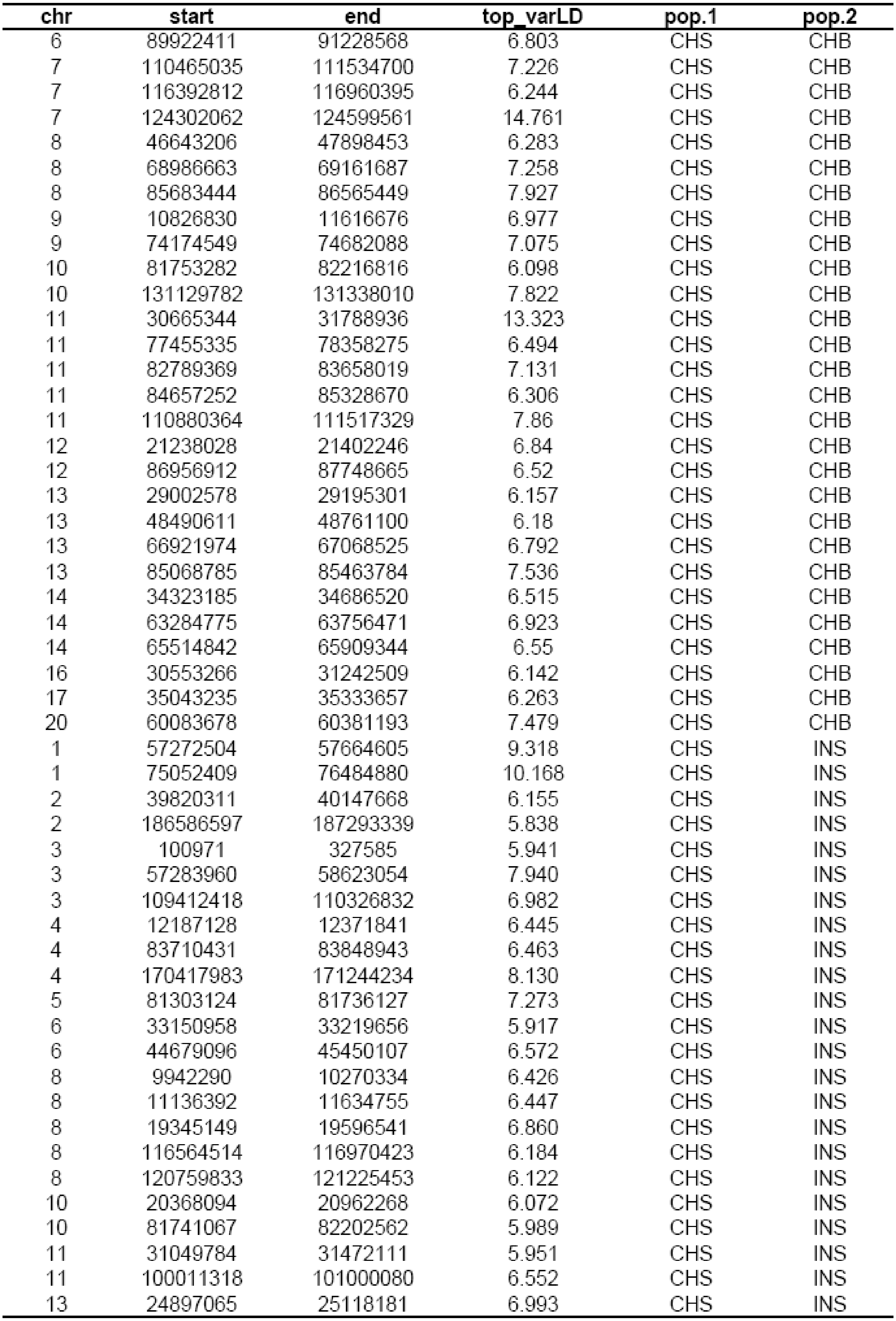
Panel	Chr: start – end (Mb, HG18)	Genes in region
CHS	1: 75.698 – 76.249	*SLC44A5*, *ACADM*, *RABGGTB*, *MSH4*, *ASB17*
CHS	3: 57.276 – 58.402	*ASB14*, *APPL1*, *2'PDE*, *ARF4*, *SLMAP*, *FLNB*, *DNASE1L3*, *ABHD6*, *RPP14*, *PXK*, *PDHB*
CHS	6: 44.861 – 45.456	*SUPT3H*, *RUNX2*
CHS	10: 81.930 – 82.199	*ANXA11*, *MAT1A*, *DYDC1*, *DYDC2*, *C10orf58*
CHS	11: 80.703 – 81.995	-
CHS	11: 83.302 – 85.530	*DLG2*, *CREBZF*, *CCDC89*, *SYTL2*, *CCDC83*, *PICALM*
CHS	11: 100.037 – 100.585	*PGR*
CHS	14: 58.830 – 59.454	*DAAM1*, *RTN1*, *GPR135*, *C14orf149*, *C14orf100*
CHS	15: 28.310 – 29.202	*CHRFAM7A*, *MTMR15*, *MTMR10*, *TRPM1*
CHS	16: 57.806 – 58.017	-
MAS	1: 45.700 – 46.404	*TESK2*, *MMACHC*, *PRDX1*, *AKR1A1*, *NASP*, *IPP*, *MAST2*, *CCDC17*, *GPBP1L1*, *TMEM69*, *PIK3R3*
MAS	1: 52.432 – 52.918	*ZFYVE9*, *CC2D1B*, *ORC1L*, *PRPF38A*, *ZCCHC11*, *GPX7*
MAS	3: 58.121 – 58.441	*FLNB*, *DNASE1L3*, *ABHD6*, *RPP14*, *PXK*, *PDHB*
MAS	8: 102.710 – 102.988	*GRHL2*, *NCALD*
MAS	10: 23.830 – 24.232	-
MAS	10: 30.331 – 30.590	-
MAS	11: 72.996 – 73.345	*PLEKHB1*, *RAB6A*, *MRPL48*, *CHCHD8*, *WDR71*
MAS	11: 83.300 – 85.536	*DLG2*, *CREBZF*, *CCDC89*, *SYTL2*, *CCDC83*, *PICALM*
MAS	12: 34.762 – 36.945	-
MAS	15: 28.353 – 29.105	*CHRFAM7A*, *MTMR15*, *MTMR10*, *TRPM1*
INS	1: 156.668 – 156.830	*OR10K1*, *OR10R2*, *OR6Y1*, *OR10X1*
INS	2: 3.092 – 5.120	*TSSC1*, *TTC15*, *ADI1*, *RNASEH1*, *RPS7*, *COLEC11*, *ALLC*
INS	5: 142.235 – 142.718	*ARHGAP26*, *NR3C1*
INS	10: 30.280 – 30.598	-
INS	11: 38.214 – 39.950	-
INS	12: 0.230 – 0.435	*SLC6A13*, *JARID1A*, *CCDC77*
INS	12: 34.866 – 37.612	*ALG10B*, *CPNE8*
INS	15: 45.795 – 46.740	*SEMA6D*, *SLC24A5*, *MYEF2*, *SLC12A1*, *DUT*, *FBN1*
INS	15: 72.382 – 72.997	*CCDC33*, *CYP11A1*, *SEMA7A*, *UBL7*, *ARID3B*, *CLK3*, *EDC3*, *CYP1A2*, *CYP1A1*, *CSK*, *LMAN1L*, *CPLX3*, *ULK3*, *SCAMP2*, *MPI*
INS	16: 57.802 – 58.173	-

Table S4. Top 0.1% candidate regions of LD variation between pairs of populations in HapMap and SGVP



chr	start	end	top_varLD	pop.1	pop.2
1	75697523	76249038	6.852	CHS	CEU
2	3918956	4888263	5.956	CHS	CEU
3	57276318	58402462	7.098	CHS	CEU
3	109440945	110352745	6.132	CHS	CEU
3	119042532	119276731	6.025	CHS	CEU
3	127640518	128251670	6.365	CHS	CEU
4	12172429	12379030	6.301	CHS	CEU
4	83660732	83850594	5.961	CHS	CEU
4	97210662	97628847	6.290	CHS	CEU
4	101498743	101583190	6.071	CHS	CEU
4	170603521	170881404	6.257	CHS	CEU
5	53330998	53578689	6.013	CHS	CEU
5	108656444	109200120	5.948	CHS	CEU
5	142213563	142432440	6.602	CHS	CEU
6	44861386	45455527	7.578	CHS	CEU
8	10611494	10786363	6.270	CHS	CEU
8	19351021	19448470	5.977	CHS	CEU
9	101767959	102397962	6.206	CHS	CEU
10	81929740	82198886	7.464	CHS	CEU
11	31070916	31523272	6.514	CHS	CEU
11	80703009	81995090	8.365	CHS	CEU
11	83302492	85530235	7.444	CHS	CEU
11	100036678	100585052	7.276	CHS	CEU
13	20734647	21128878	6.059	CHS	CEU
14	58830298	59454154	7.128	CHS	CEU
14	69236368	69703897	6.032	CHS	CEU
15	28310071	29202639	8.911	CHS	CEU
16	57805588	58017313	7.447	CHS	CEU
20	1177896	1640634	6.483	CHS	CEU
20	27183146	29906510	6.410	CHS	CEU
1	45035456	46208579	8.282	CHS	CHB
1	202744727	204245628	6.429	CHS	CHB
2	3046534	5049101	13.331	CHS	CHB
2	14725607	14944696	7.888	CHS	CHB
3	33460287	33936985	7.776	CHS	CHB
3	34617647	34797744	9.804	CHS	CHB
3	41092294	41671758	6.133	CHS	CHB
3	57315131	57559223	6.148	CHS	CHB
3	120833764	122069288	6.401	CHS	CHB
3	126825131	128123418	7.214	CHS	CHB
3	181310003	182158580	7.505	CHS	CHB
4	50424593	52832648	7.223	CHS	CHB
4	96700979	97087972	6.25	CHS	CHB
5	13393939	13943858	6.814	CHS	CHB
5	71599103	73532763	8.017	CHS	CHB
5	100634247	100928800	7.503	CHS	CHB
5	134740058	135151101	6.715	CHS	CHB
6	29603413	31455448	8.462	CHS	CHB
6	32338573	33488580	12.476	CHS	CHB
6	78071210	79262867	6.583	CHS	CHB
6	83089440	84829648	7.578	CHS	CHB

Table S4. Continued

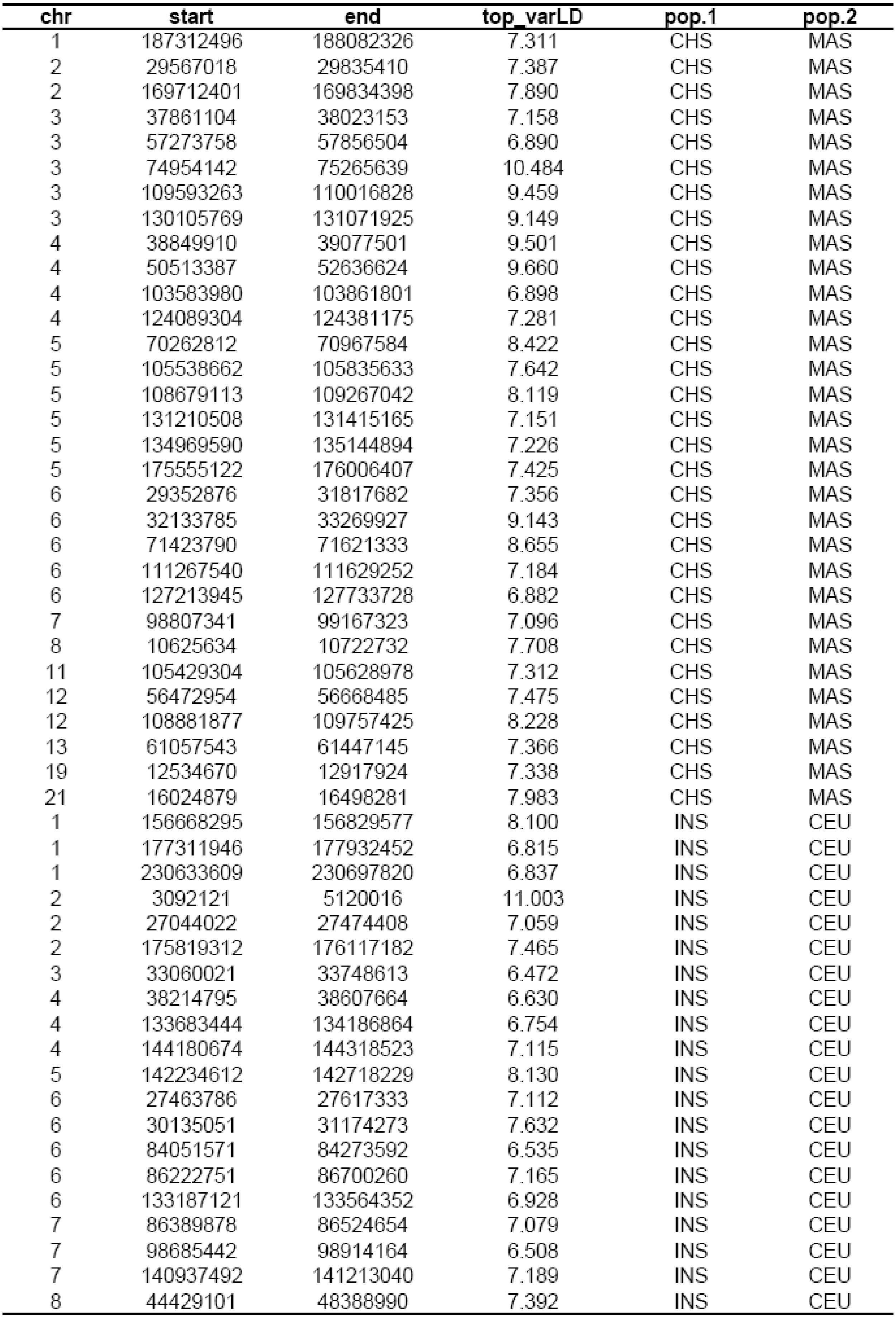


chr	start	end	top_varLD	pop.1	pop.2
6	89922411	91228568	6.803	CHS	CHB
7	110465035	111534700	7.226	CHS	CHB
7	116392812	116960395	6.244	CHS	CHB
7	124302062	124599561	14.761	CHS	CHB
8	46643206	47898453	6.283	CHS	CHB
8	68986663	69161687	7.258	CHS	CHB
8	85683444	86565449	7.927	CHS	CHB
9	10826830	11616676	6.977	CHS	CHB
9	74174549	74682088	7.075	CHS	CHB
10	81753282	82216816	6.098	CHS	CHB
10	131129782	131338010	7.822	CHS	CHB
11	30665344	31788936	13.323	CHS	CHB
11	77455335	78358275	6.494	CHS	CHB
11	82789369	83658019	7.131	CHS	CHB
11	84657252	85328670	6.306	CHS	CHB
11	110880364	111517329	7.86	CHS	CHB
12	21238028	21402246	6.84	CHS	CHB
12	86956912	87748665	6.52	CHS	CHB
13	29002578	29195301	6.157	CHS	CHB
13	48490611	48761100	6.18	CHS	CHB
13	66921974	67068525	6.792	CHS	CHB
13	85068785	85463784	7.536	CHS	CHB
14	34323185	34686520	6.515	CHS	CHB
14	63284775	63756471	6.923	CHS	CHB
14	65514842	65909344	6.55	CHS	CHB
16	30553266	31242509	6.142	CHS	CHB
17	35043235	35333657	6.263	CHS	CHB
20	60083678	60381193	7.479	CHS	CHB
1	57272504	57664605	9.318	CHS	INS
1	75052409	76484880	10.168	CHS	INS
2	39820311	40147668	6.155	CHS	INS
2	186586597	187293339	5.838	CHS	INS
3	100971	327585	5.941	CHS	INS
3	57283960	58623054	7.940	CHS	INS
3	109412418	110326832	6.982	CHS	INS
4	12187128	12371841	6.445	CHS	INS
4	83710431	83848943	6.463	CHS	INS
4	170417983	171244234	8.130	CHS	INS
5	81303124	81736127	7.273	CHS	INS
6	33150958	33219656	5.917	CHS	INS
6	44679096	45450107	6.572	CHS	INS
8	9942290	10270334	6.426	CHS	INS
8	11136392	11634755	6.447	CHS	INS
8	19345149	19596541	6.860	CHS	INS
8	116564514	116970423	6.184	CHS	INS
8	120759833	121225453	6.122	CHS	INS
10	20368094	20962268	6.072	CHS	INS
10	81741067	82202562	5.989	CHS	INS
11	31049784	31472111	5.951	CHS	INS
11	100011318	101000080	6.552	CHS	INS
13	24897065	25118181	6.993	CHS	INS

Table S4. Continued

chr	start	end	top_varLD	pop.1	pop.2
13	62159775	62678934	6.177	CHS	INS
13	78754699	79002435	6.521	CHS	INS
15	28640049	29218421	8.059	CHS	INS
15	72510001	72725772	7.570	CHS	INS
17	21998424	22897226	6.699	CHS	INS
17	25017632	25540026	6.665	CHS	INS
18	16339945	17506475	6.814	CHS	INS
20	34929286	35275164	7.877	CHS	INS
21	33601836	33884274	5.967	CHS	INS
22	21765940	21900600	6.857	CHS	INS
1	45611533	46299926	8.043	CHS	JPT+CHB
1	77799045	78077082	7.686	CHS	JPT+CHB
1	148767228	150433794	7.702	CHS	JPT+CHB
1	156628841	156857342	10.307	CHS	JPT+CHB
2	3037784	5149038	16.859	CHS	JPT+CHB
2	14749560	14915469	7.438	CHS	JPT+CHB
2	135531371	135706027	6.958	CHS	JPT+CHB
3	8897477	8979272	7.683	CHS	JPT+CHB
3	72142870	72232709	6.919	CHS	JPT+CHB
3	75023161	75282518	11.545	CHS	JPT+CHB
3	127029073	127966367	7.915	CHS	JPT+CHB
3	181310003	182208777	8.477	CHS	JPT+CHB
4	120322951	120753736	10.071	CHS	JPT+CHB
4	124175294	124438834	8.977	CHS	JPT+CHB
5	13696256	13766374	7.322	CHS	JPT+CHB
5	128892752	129305077	7.158	CHS	JPT+CHB
5	134969866	135057316	7.271	CHS	JPT+CHB
6	29728376	31457421	6.959	CHS	JPT+CHB
6	32801118	33478103	12.303	CHS	JPT+CHB
6	85060723	86573873	7.093	CHS	JPT+CHB
8	68832563	69180640	7.732	CHS	JPT+CHB
8	71240836	72054691	7.159	CHS	JPT+CHB
9	74153815	74726549	8.495	CHS	JPT+CHB
10	64524644	64906472	11.585	CHS	JPT+CHB
11	30988484	31558901	6.966	CHS	JPT+CHB
11	82760973	82986237	9.407	CHS	JPT+CHB
11	91499873	92126456	8.364	CHS	JPT+CHB
12	21318390	21383794	11.032	CHS	JPT+CHB
13	63589236	63741624	7.690	CHS	JPT+CHB
14	55471902	56108438	8.521	CHS	JPT+CHB
14	63224333	63601418	7.562	CHS	JPT+CHB
15	46338987	46672595	8.328	CHS	JPT+CHB
15	57803715	57914950	7.882	CHS	JPT+CHB
20	53844014	53944689	8.191	CHS	JPT+CHB
22	37258390	37465125	8.763	CHS	JPT+CHB
1	45971103	46339621	13.144	CHS	MAS
1	57551420	57868776	11.094	CHS	MAS
1	75522552	76267988	9.539	CHS	MAS
1	106635331	107205422	7.213	CHS	MAS
1	150001597	151032776	8.785	CHS	MAS
1	153294770	154120104	11.073	CHS	MAS

Table S4. Continued

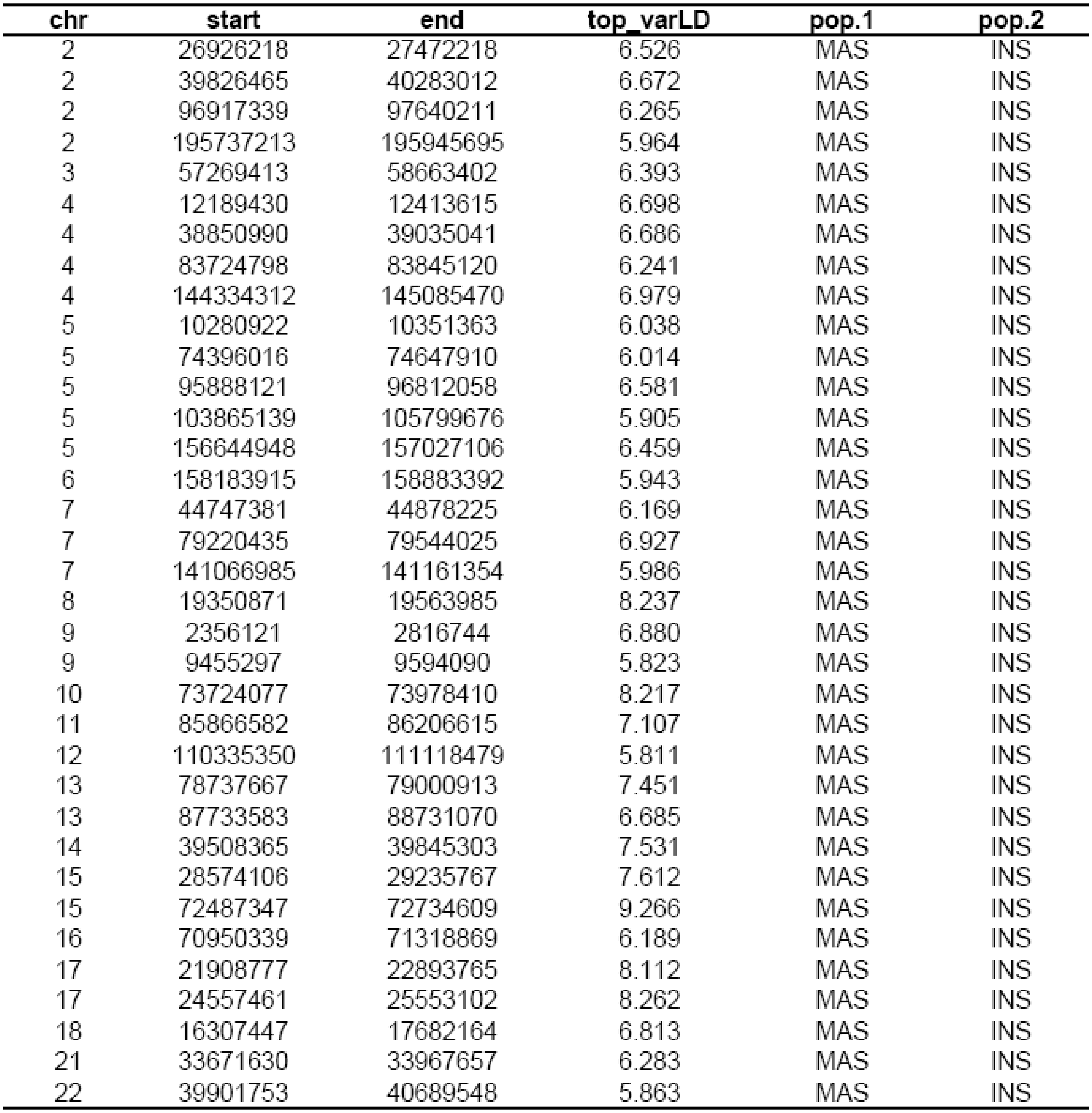


chr	start	end	top_varLD	pop.1	pop.2
1	187312496	188082326	7.311	CHS	MAS
2	29567018	29835410	7.387	CHS	MAS
2	169712401	169834398	7.890	CHS	MAS
3	37861104	38023153	7.158	CHS	MAS
3	57273758	57856504	6.890	CHS	MAS
3	74954142	75265639	10.484	CHS	MAS
3	109593263	110016828	9.459	CHS	MAS
3	130105769	131071925	9.149	CHS	MAS
4	38849910	39077501	9.501	CHS	MAS
4	50513387	52636624	9.660	CHS	MAS
4	103583980	103861801	6.898	CHS	MAS
4	124089304	124381175	7.281	CHS	MAS
5	70262812	70967584	8.422	CHS	MAS
5	105538662	105835633	7.642	CHS	MAS
5	108679113	109267042	8.119	CHS	MAS
5	131210508	131415165	7.151	CHS	MAS
5	134969590	135144894	7.226	CHS	MAS
5	175555122	176006407	7.425	CHS	MAS
6	29352876	31817682	7.356	CHS	MAS
6	32133785	33269927	9.143	CHS	MAS
6	71423790	71621333	8.655	CHS	MAS
6	111267540	111629252	7.184	CHS	MAS
6	127213945	127733728	6.882	CHS	MAS
7	98807341	99167323	7.096	CHS	MAS
8	10625634	10722732	7.708	CHS	MAS
11	105429304	105628978	7.312	CHS	MAS
12	56472954	56668485	7.475	CHS	MAS
12	108881877	109757425	8.228	CHS	MAS
13	61057543	61447145	7.366	CHS	MAS
19	12534670	12917924	7.338	CHS	MAS
21	16024879	16498281	7.983	CHS	MAS
1	156668295	156829577	8.100	INS	CEU
1	177311946	177932452	6.815	INS	CEU
1	230633609	230697820	6.837	INS	CEU
2	3092121	5120016	11.003	INS	CEU
2	27044022	27474408	7.059	INS	CEU
2	175819312	176117182	7.465	INS	CEU
3	33060021	33748613	6.472	INS	CEU
4	38214795	38607664	6.630	INS	CEU
4	133683444	134186864	6.754	INS	CEU
4	144180674	144318523	7.115	INS	CEU
5	142234612	142718229	8.130	INS	CEU
6	27463786	27617333	7.112	INS	CEU
6	30135051	31174273	7.632	INS	CEU
6	84051571	84273592	6.535	INS	CEU
6	86222751	86700260	7.165	INS	CEU
6	133187121	133564352	6.928	INS	CEU
7	86389878	86524654	7.079	INS	CEU
7	98685442	98914164	6.508	INS	CEU
7	140937492	141213040	7.189	INS	CEU
8	44429101	48388990	7.392	INS	CEU

Table S4. Continued

chr	start	end	top_varLD	pop.1	pop.2
8	64685250	65185054	6.465	INS	CEU
10	23838197	24258291	7.391	INS	CEU
10	30279666	30598488	7.882	INS	CEU
11	38214296	39950217	8.290	INS	CEU
12	229569	434908	8.297	INS	CEU
12	34865673	37611921	8.344	INS	CEU
13	78753242	79213288	6.809	INS	CEU
14	51428246	51787530	6.426	INS	CEU
14	58856702	59479139	7.483	INS	CEU
15	45795006	46740464	7.807	INS	CEU
15	72381699	72996706	8.276	INS	CEU
15	80234159	81318484	7.252	INS	CEU
16	57802382	58172878	8.953	INS	CEU
17	25139410	25519301	7.131	INS	CEU
18	61613394	62356729	6.567	INS	CEU
1	45699517	46404274	8.464	MAS	CEU
1	52431990	52917759	7.577	MAS	CEU
1	147612020	148851232	6.227	MAS	CEU
1	156644164	156823261	6.741	MAS	CEU
1	177309239	178049429	6.617	MAS	CEU
2	162909499	163552763	6.707	MAS	CEU
3	58120541	58441220	8.410	MAS	CEU
4	12187913	12432814	6.407	MAS	CEU
4	133773937	134258406	6.876	MAS	CEU
4	144806224	145120652	6.409	MAS	CEU
5	138186452	138682656	6.603	MAS	CEU
6	116621860	117019116	6.282	MAS	CEU
7	83896426	84503657	6.299	MAS	CEU
7	124410311	124600987	6.265	MAS	CEU
8	102709578	102988494	8.285	MAS	CEU
9	101753392	102393693	6.545	MAS	CEU
10	23829866	24232288	7.556	MAS	CEU
10	30331078	30589833	7.045	MAS	CEU
10	49862335	50062352	6.309	MAS	CEU
10	60520733	60648766	6.658	MAS	CEU
10	61173330	61380843	6.348	MAS	CEU
10	131135331	131324954	6.713	MAS	CEU
11	72996225	73344624	6.896	MAS	CEU
11	83299618	85535632	8.344	MAS	CEU
12	34762302	36944777	8.221	MAS	CEU
12	45030368	45399088	6.479	MAS	CEU
12	110294798	111146154	6.826	MAS	CEU
13	81376922	81977299	6.344	MAS	CEU
13	87874729	88378323	6.896	MAS	CEU
14	58841046	59486284	6.871	MAS	CEU
14	76835247	77341335	6.306	MAS	CEU
15	28352789	29105127	8.080	MAS	CEU
15	46142608	46712775	6.804	MAS	CEU
20	1304056	1622736	6.469	MAS	CEU
1	52431833	52920666	10.129	MAS	INS
1	96113119	96744891	6.500	MAS	INS

Table S4. Continued



chr	start	end	top_varLD	pop.1	pop.2
2	26926218	27472218	6.526	MAS	INS
2	39826465	40283012	6.672	MAS	INS
2	96917339	97640211	6.265	MAS	INS
2	195737213	195945695	5.964	MAS	INS
3	57269413	58663402	6.393	MAS	INS
4	12189430	12413615	6.698	MAS	INS
4	38850990	39035041	6.686	MAS	INS
4	83724798	83845120	6.241	MAS	INS
4	144334312	145085470	6.979	MAS	INS
5	10280922	10351363	6.038	MAS	INS
5	74396016	74647910	6.014	MAS	INS
5	95888121	96812058	6.581	MAS	INS
5	103865139	105799676	5.905	MAS	INS
5	156644948	157027106	6.459	MAS	INS
6	158183915	158883392	5.943	MAS	INS
7	44747381	44878225	6.169	MAS	INS
7	79220435	79544025	6.927	MAS	INS
7	141066985	141161354	5.986	MAS	INS
8	19350871	19563985	8.237	MAS	INS
9	2356121	2816744	6.880	MAS	INS
9	9455297	9594090	5.823	MAS	INS
10	73724077	73978410	8.217	MAS	INS
11	85866582	86206615	7.107	MAS	INS
12	110335350	111118479	5.811	MAS	INS
13	78737667	79000913	7.451	MAS	INS
13	87733583	88731070	6.685	MAS	INS
14	39508365	39845303	7.531	MAS	INS
15	28574106	29235767	7.612	MAS	INS
15	72487347	72734609	9.266	MAS	INS
16	70950339	71318869	6.189	MAS	INS
17	21908777	22893765	8.112	MAS	INS
17	24557461	25553102	8.262	MAS	INS
18	16307447	17682164	6.813	MAS	INS
21	33671630	33967657	6.283	MAS	INS
22	39901753	40689548	5.863	MAS	INS

The top_varLD column contains the maximum standardized score found within the start and end position of each region. This list contains only the regions in the top 0.1% of the varLD score distribution for each population pair.

Table S5. Top 10 candidate regions of LD variation between CHS and CHB.

Chr: start – end (Mb, HG18)	Genes in region
1: 156.629 – 156.857	*OR10T2*, *OR10K2*, *OR10K1*, *OR10R2*, *OR6Y1*, *OR10X1*, *OR10Z1*, *SPTA1*
2: 3.038 – 5.149	*TSSC1*, *TTC15*, *ADI1*, *RNASEH1*, *RPS7*, *COLEC11*, *ALLC*
3: 75.023 – 75.283	-
4: 120.323 – 120.754	*MYOZ2*, *USP53*, *FABP2*, *PDE5A*
4: 124.175 – 124.439	*SPATA5*
6: 32.801 – 33.478	*HLA-DQA/B* gene clusters, *TAP2*, *PSMB9*, *TAP1*, *BRD2*, *PSMB8*, *COL11A2*, *RPL32P1*, *VPS52*, *B3GALT*, *TAPBP*, *RXRB*, *RING1*, *DAXX*, *WDR46*, *PFDN6*, *RPS18*, *RGL2*, *ZBTB22*, *ZNF314P*
10: 64.525 – 64.906	*NRBF2*, *JMJD1C*
11: 82.761 – 82.986	*DLG2*
12: 21.318 – 21.384	*SLCO1A2*
22: 37.258 – 37.465	*DMC1*, *CBY1*, *TOMM22*, *JOSD1*, *GTPBP1*, *UNC84B*

Table S6. Breakdown of the regions containing putative signals of positive selection identified by iHS

	Number of SNPs with standardized |iHS| > 2*				
Population	**3 – 5**	**6 – 10**	**11 – 20**	**> 20**	**Total**
CHS	1988	1018	607	412	4025
in HapMap	1623	920	568	389	3499
novel (common**)	204	62	30	15	311
novel (unique)	161	37	9	8	215
MAS	1959	1063	592	426	4040
in HapMap	1567	933	550	403	3453
novel (common)	205	85	25	17	332
novel (unique)	187	45	17	6	255
INS	2025	1103	632	454	4214
in HapMap	1626	973	575	417	3591
novel (common)	128	59	29	21	237
novel (unique)	271	71	28	16	386

* Regions containing less than 3 SNPs are excluded from the analyses.

** A signal in the same region is observed in at least one of the other two SGVP populations.

Table S7. Top 10 candidate regions for recent positive natural selection by iHS in each of the SGVP populations, and whether it was previously observed in any of the HapMap panels

Chr	Bin start	Bin end	Genes in region	Peak SNP	HapMap
CHS					
2	17,400,000	17,800,000	*VSNL1*, *SMC6*, *GEN1*	rs2344691	No
2	25,800,000	26,400,000	*ASXL2*, *KIF3C*, *HADHA*, *RAB10*, *HADHB*, *GPR113*	rs11685550	No
2	108,300,000	108,500,000	*SULT1C2*, *GCC2*	rs10169264	Yes
2	125,200,000	126,100,000	*CNTNAP5*	rs9308661	No
2	197,300,000	197,500,000	*GTF3C3*, *PGAP1*	rs16857456	Yes
3	108,900,000	109,200,000	*BBX*	rs1437240	No
4	143,700,000	144,600,000	*USP38*, *GAB1*	rs12501994	Yes
7	5,400,000	5,800,000	*KIAA1856*, *FBXL18*, *ACTB*, *FSCN1*, *TRIAD3*	rs852441	No
10	107,200,000	107,500,000	-	rs7091254	Yes
12	1,100,000	1,400,000	*ERC1*	rs2286031	No
MAS					
1	153,100,000	153,400,000	*PMVK*, *PBXIP1*, *PYGO2*, *SHC1*, *CKS1B*, *FLAD1*, *LENEP*, *ZBTB7B*, *DCST2*, *ADAM15*, *DCST1*, *EFNA4*, *EFNA3*, *EFNA1*, *RAG1AP1*, *DPM3*	rs4845681	No
2	84,300,000	84,900,000	*SUCLG1*	rs1192368	No
3	108,600,000	109,200,000	*BBX*	rs329921	No
5	117,400,000	117,900,000	-	rs11743225	No
8	67,000,000	67,100,000	-	rs435575	No
10	94,400,000	95,100,000	*KIF11*, *HHEX*, *EXOC6*, *CYP26A1*, *CYP26C1*, *FER1L3*	rs7091432	Yes
11	25,100,000	25,600,000	-	rs2404091	Yes
12	87,000,000	87,600,000	*CEP290*, *TMTC3*, *KITLG*	rs1508595	No
15	61,600,000	62,600,000	*USP3*, *FBXL22*, *HERC1*, *DAPK2*, *FAM96A*, *SNX1*, *SNX22*, *PPIB*, *CSNK1G1*, *TRIP4*, *ZNF609*	rs16947748	Yes
17	25,000,000	25,500,000	*SSH2*, *EFCAB5*, *CCDC55*	rs7226121	No
INS					
2	82,800,000	83,100,000	-	rs897383	No
2	96,300,000	97,100,000	*SNRNP200*, *NCAPH*, *ITRIPL1*, *NEURL3*, *ARID5A*, *FER1L5*, *CNNM4*, *CNNM3*, *SEMA4C*, *ANKRD23*, *ANKRD39*, *FAM178B*	rs17420101	No
4	29,100,000	30,000,000	-	rs11722527	No
4	32,900,000	34,200,000	-	rs10517297	Yes
4	41,500,000	41,900,000	*TMEM33*, *WDR21B*, *SLC30A9*, *CCDC4*	rs2343617	Yes
7	119,500,000	120,300,000	*KCND2*, *TSPAN12*	rs4730954	No
8	42,600,000	42,800,000	*CHRNB3*, *CHRNA6*	rs11986893	No
11	60,600,000	61,000,000	*CD5*, *VPS37C*, *PGA3*, *PGA4*, *PGA5*, *VWCE*, *DOB1*, *DAK*, *CYBASC3*, *FLJ12529*, *C11orf79*	rs3019198	No
16	30,800,000	31,100,000	*CTF1*, *FBXL19*, *ORAI3*, *SETD1A*, *STX4*, *BCKDK*, *HSD3B7*, *STX1B2*, *ZNF668*, *ZNF646*, *VKORC1*, *PRSS8*, *TRIM72*, *PRSS36*, *MYST1*, *FUS*	rs17839567	No
17	24,900,000	25,900,000	*TP53I13*, *GIT1*, *ANKRD13B*, *CORO6*, *SSH2*, *EFCAB5*, *CCDC55*, *SLC6A4*, *BLMH*, *TMIGD1*, *CPD*, *GOSR1*	rs10445400	No

Table S8. Signals of positive natural selection for regions that are discussed in the main text and Table S7.

Population	rsID	chr	start	end	snps_ihs	top_ihs	snps_xpehh	top_xpehh	HapMap	Gene
CHS	rs9439603	1	65499104	65939087	46	3.25	140	3.70	Yes	LEPR
CHS	rs10519439	2	21165196	21452155	28	3.46	1	2.65	Yes	APOB
CHS	rs17210194	3	166216415	166294201	17	2.69	0	---	Yes	SI
CHS	rs10212960	4	99678000	100706735	63	3.69	178	3.92	Yes	ADH cluster
CHS	**rs755447**	**4**	**120600789**	**120617277**	**5**	**2.25**	**15**	**2.67**	**No**	**FABP2**
CHS	**rs6979074**	**7**	**64881031**	**65112236**	**7**	**2.05**	**7**	**2.98**	**Yes**	**VKORC1**
CHS	rs10813630	9	12415349	12642983	21	3.04	12	2.75	Yes	TYRP1
CHS	rs657391	9	120198462	120493431	13	2.69	0	---	Yes	CDK5RAP2
CHS	rs9511107	13	24087224	24407384	19	2.59	0	---	Yes	CENPJ
CHS	rs2078094	15	25676930	25984569	18	2.89	75	6.02	Yes	OCA2
CHS	rs11161121	15	50230366	50500294	13	2.99	0	---	Yes	MYO5A
CHS	rs7249235	19	11103133	11103765	3	2.58	1	2.52	Yes	LDLR
INS	rs12714396	2	21173986	21431248	39	3.22	0	---	Yes	APOB
INS	rs4499656	4	100469993	100493309	9	2.55	24	4.25	Yes	ADH cluster
INS	rs1345196	8	6158400	6361761	10	2.48	2	2.96	Yes	MCHP1
INS	rs10970464	9	12711806	12753450	6	2.45	7	2.80	Yes	TYRP1
INS	rs7040388	9	120413639	120463654	9	3.08	0	---	Yes	CDK5RAP2
INS	rs17071834	13	24259850	24417796	8	2.36	0	---	Yes	CENPJ
INS	**rs8030283**	**15**	**46163796**	**46752510**	**121**	**4.18**	**270**	**5.35**	**Yes**	**SLC24A5**
INS	rs12442023	15	50500294	50501372	3	2.59	4	2.96	Yes	MYO5A
INS	**rs8051399**	**16**	**30730548**	**31187037**	**48**	**3.78**	**0**	**---**	**Yes**	**VKORC1**
INS	rs732310	19	11193505	11219440	4	2.39	4	2.56	Yes	LDLR
MAS	rs10753360	1	65616241	66353820	54	3.18	84	3.11	Yes	LEPR
MAS	rs10865547	2	21165196	21327361	12	2.57	0	---	Yes	APOB
MAS	rs6442317	3	166216415	166294201	13	2.72	0	---	Yes	SI
MAS	rs7460646	8	6327592	6430788	4	2.23	28	3.13	Yes	MCPH1
MAS	rs2022011	9	120413639	120463654	8	2.84	0	---	Yes	CDK5RAP2
MAS	rs8029455	15	25676930	25781616	4	2.35	0	---	Yes	OCA2
MAS	rs2353506	15	50487842	50501372	4	3.08	0	---	Yes	MYO5A

Regions that emerged from F_{ST} scans and surveys of LD variations are highlighted in bold. Number of SNPs with |iHS| > 2 and XP-EHH scores > 2.5 are shown in each candidate region. The highest score observed across the identified SNPs for each of the two methods is also shown. We also denote whether the selection signal was previously observed in the HapMap.

Table S9. Extent of IBS for Affymetrix samples

Sample ID 1	Sample ID 2	Missingness Sample 1 (%)	Missingness Sample 2 (%)	Similarity (%)	Possible relationship	Sample ID removed
002_1	002_2	0.82	0.15	99.919	Duplicate	002_1
063_1	063_2	2.47	0.12	99.417	Duplicate	063_1
250_1	250_2	1.51	0.47	99.676	Duplicate	250_1
446_1	444_1	0.65	0.25	88.344	Siblings	446_1
314_1	329_1	0.33	0.65	87.147	Siblings	329_1
168_1	398_1	0.30	0.36	87.383	Siblings	398_1
442_1	472_1	1.10	0.51	88.030	Siblings	442_1
500_1	518_1	0.65	1.20	87.480	Siblings	500_1 518_1
500_1	516_1	0.65	0.19	88.656	Siblings	
516_1	518_1	0.19	1.20	87.623	Siblings	

Table S10. Extent of IBS for Illumina samples

Sample ID 1	Sample ID 2	Missingness Sample 1 (%)	Missingness Sample 2 (%)	IBS (%)	Possible relationship	Sample ID removed
002_1	002_2	0.25	0.08	99.994	Duplicate	002_1
063_1	063_2	0.34	0.20	99.974	Duplicate	063_1
250_1	250_2	0.27	0.15	99.992	Duplicate	250_1
406_1	492_1	0.06	0.06	100.000	Duplicate	406_1
446_1	444_1	1.28	0.05	76.854	Siblings	446_1
314_1	329_1	0.21	0.08	74.173	Siblings	314_1
168_1	398_1	11.19	0.16	74.684	Siblings	168_1
442_1	472_1	0.06	0.11	76.154	Siblings	472_1
500_1	518_1	0.14	0.11	74.993	Siblings	500_1 516_1
500_1	516_1	0.14	0.12	77.057	Siblings	
516_1	518_1	0.11	0.12	75.327	Siblings	

Table S11. Samples with discordant self-reported and PCA-inferred population membership.

Sample ID	Reported population	PCA inferred population membership
323_1	Indian	Possible admixed between Chinese and Indian
137_1	Chinese	Possible admixed between Chinese and Malay
383_1	Indian	Possible Malay membership
086_1	Indian	Possible admixed between Indian and Malay
495_1	Indian	Possible Malay membership
267_1	Indian	Possible admixed between Chinese, Malay and Indian
194_1	Malay	Possible Chinese membership

Table S12. SNPs with strand synchronization issues

snp-id	similarity	Flipped similarity	N	Illumina				Affymetrix				Illumina Alleles	Affymetrix Alleles
				0A	1A	2A	-1A	0B	1B	2B	-1B		
rs16942821	0	1	267	268	0	0	0	0	0	267	1	C/T	A/C
rs238137	0.007463	0.865672	268	0	0	268	0	232	34	2	0	C/T	C/T
rs348238	0	1	267	268	0	0	0	0	0	267	1	A/C	A/C
rs624307	0	1	268	268	0	0	0	0	0	268	0	C/T	C/T
rs7299820	0	0.988764	267	0	0	267	1	265	3	0	0	C/T	C/T
rs11054689	0.065134	0.678161	261	268	0	0	0	17	67	177	7	A/G	A/G
rs12991373	0.346008	0.152091	263	0	0	266	2	41	132	92	3	A/G	A/G
rs16872571	0.548507	0.298507	268	80	116	72	0	164	93	11	0	C/T	C/T
rs17151531	0.363296	0.161049	267	0	1	267	0	42	129	96	1	C/T	C/T
rs2435044	0.029851	0.660448	268	0	0	268	0	177	83	8	0	A/C	A/C
rs6119075	0.109023	0.406015	266	0	0	267	1	109	129	29	1	A/G	A/G
rs7406414	0.079245	0.550943	265	0	0	266	2	148	98	21	1	A/G	A/G

Table S13. SNPs probing different alleles when mapped to the forward strand

SNP	**Illumina Hap1M Alleles**	**Affymetrix SNP6 Alleles**	**dbSNP polymorphism**
rs7171243	C/T	C/G	C/G/T
rs16942821	C/T	A/C	A/C/T

Table S14. Number of SNPs identified for exclusion based on the extent of concordance of genotypes for SNPs common to both Affymetrix6.0 and Illumina1M

		Number of SNPs to remove	
Concordance	**Chinese**	**Malay**	**Indian**
< 0.99	16905	18289	18242
< 0.97	1251	1377	1366
< 0.95	482	552	536
< 0.90	191	244	239
< 0.85	133	159	152
< 0.80	94	116	106

Table S15. Genotyping accuracy for SNPs common to both platforms.

	Chinese	**Malay**	**Indian**	**Combined**
Concordance	99.911%	99.897%	99.888%	99.899%
Call rates	99.340%	99.263%	99.245%	99.285%

7 Data available for download and browsing

The web resource for the Singapore Genome Variation Project is hosted at: http://www.nus-cme.org.sg/SGVP/.

The data setup for bulk download only provides the genotype data for all samples after QC. The raw unfiltered data for each of Affymetrix and Illumina array are available by request to cmetyy@nus.edu.sg, cmesx@nus.edu.sg or ephcks@nus.edu.sg.

Post-QC genotype data

Genotype data for each individual that passes QC are available for download. These are agglomerated across both Affymetrix and Illumina platforms, and consist of only SNPs that pass QC. The file format consists of one file per chromosome per population. Allele designations have been mapped to the positive strand.

Frequencies

Allele and genotype frequencies for the post-QC SNPs calculated for samples that pass QC are available. The file format consists of one file per chromosome per population. The allele designations have been mapped to the positive strand.

LD data

Linkage disequilibrium summaries for SNPs found within 250kb of each other are available. The file format consists of one file per chromosome per population.

Phased haplotypes

Phased haplotypes for post-QC SNPs are generated using *fastPHASE* for post-QC SNPs. The file format consists of one file per chromosome per population in the 0/1 format with a corresponding legend file defining the allele designation.

Recombination rates

Recombination rates and genetic distances for each population are estimated using *LDhat*, and consists of one file per chromosome per group.

varLD scores for LD variation

Genome-wide varLD scores are available for 17 population-pairs, consisting of: (1) CEU-CHS; (2) CEU-JPT+CHB; (3) CEU-INS; (4) CEU-MAS; (5) CEU-YRI; (6) CHB-CHS; (7) CHS-INS; (8) CHS-JPT; (9) CHS-JPT+CHB; (10) CHS-MAS; (11) CHS-YRI; (12) INS-JPT+CHB; (13) INS-MAS; (14) INS-YRI; (15) JPT+CHB-MAS; (16) JPT+CHB-YRI; (17) MAS-YRI. The file format consists of one file per chromosome for each population-pair.

iHS scores for positive selection

The iHS score for each individual SNP is available for each population. The file format consists of one file per chromosome per group.

Genome browser

To facilitate the download and browsing of the SGVP data, we have designed a web-browser modeled after the version provided by the HapMap and the Human Genome Diversity Project. This uses the source codes for the Generic Genome Browser version 1.68 [11], with modifications to display the results from varLD and iHS. This can be accessed from http://www.nus-cme.org.sg/SGVP/.

8 References

1. Korn,JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40: 1253–1260.
2. Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. (2002) BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques 2002 Suppl: 56–58, 60–61.
3. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, et al. (2004) High Parallel SNP Genotyping. Cold Spring Harb Symp Quant Biol 68: 69–78.
4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
5. The International HapMap Consortium. (2005) A haplotype map of the human genome. Nature 427: 1299–1320.
6. Servin B, Stephens M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3: e114.
7. Barrett JC, Fry B, Maller J, Daly MJ. (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
8. Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, Clark TG. (2009) Genome-wide comparisons of variation in linkage disequilibrium. Genome Res [epub 18th June 2009].
9. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
10. Voight BF, Kudaravalli S, Wen X, Pritchard JK. (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
11. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, et al. (2002) The generic genome browser: a building block for a model organism system database. Genome Res 12:1599–1610.