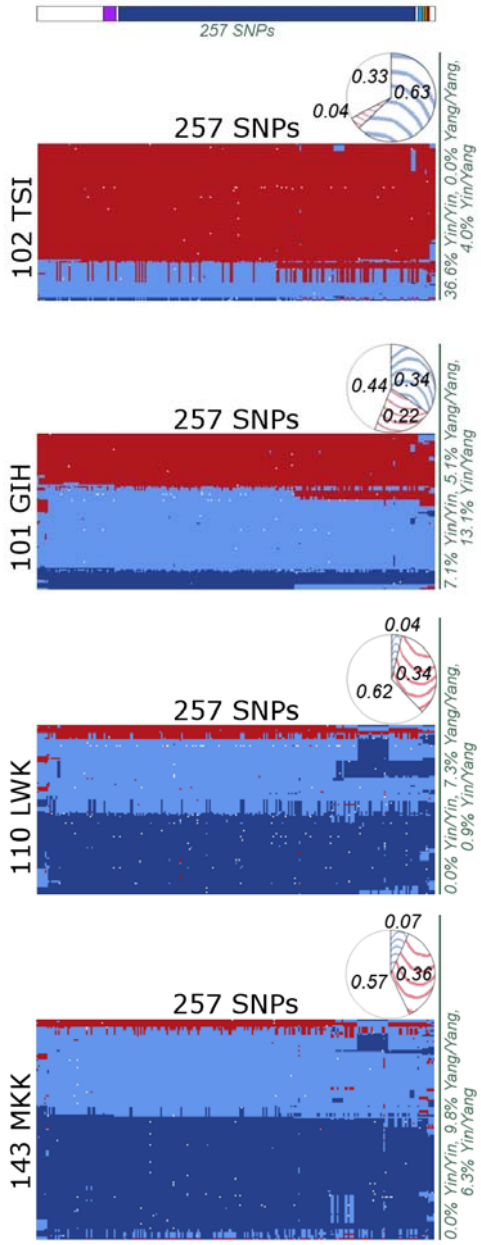
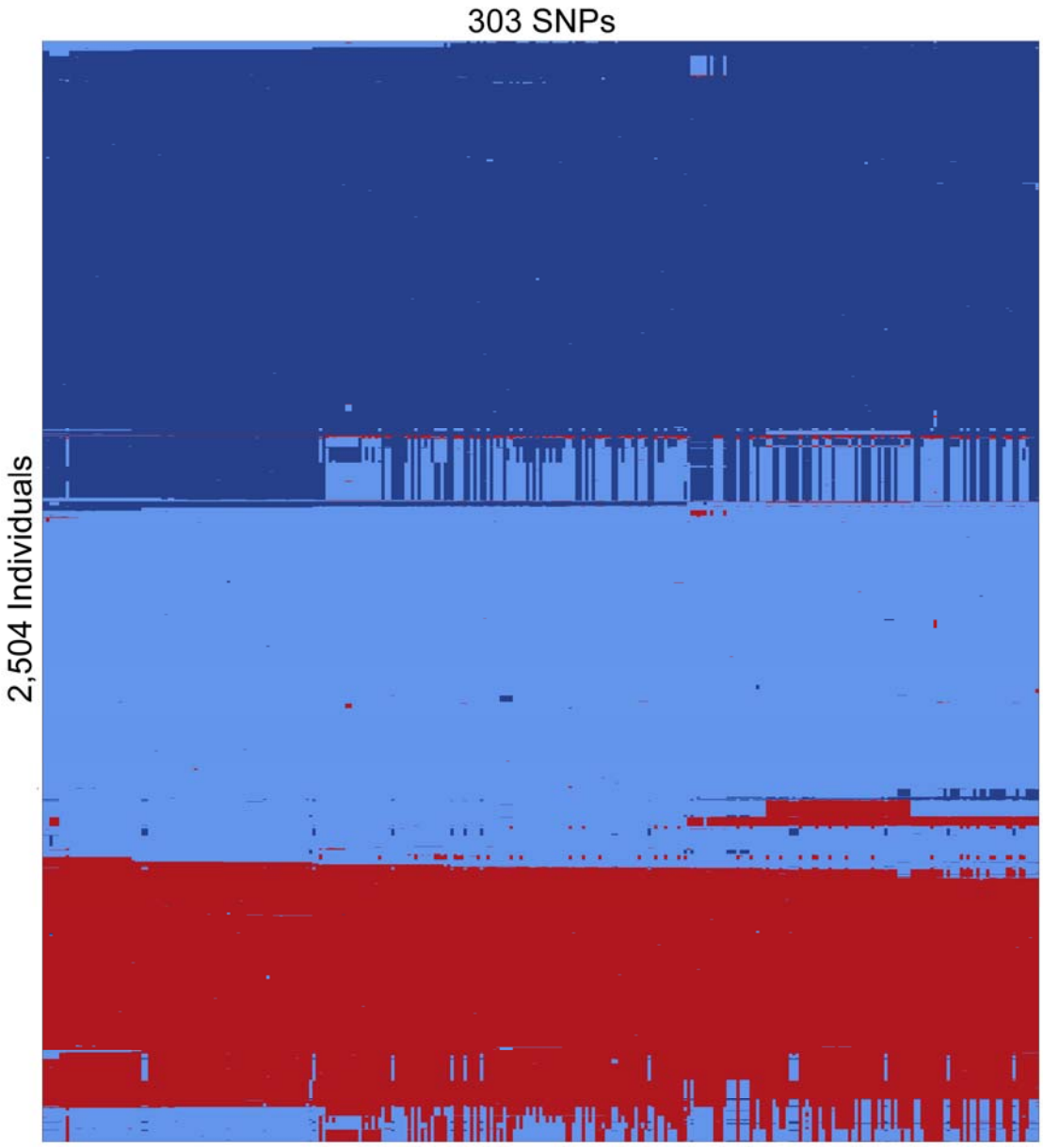
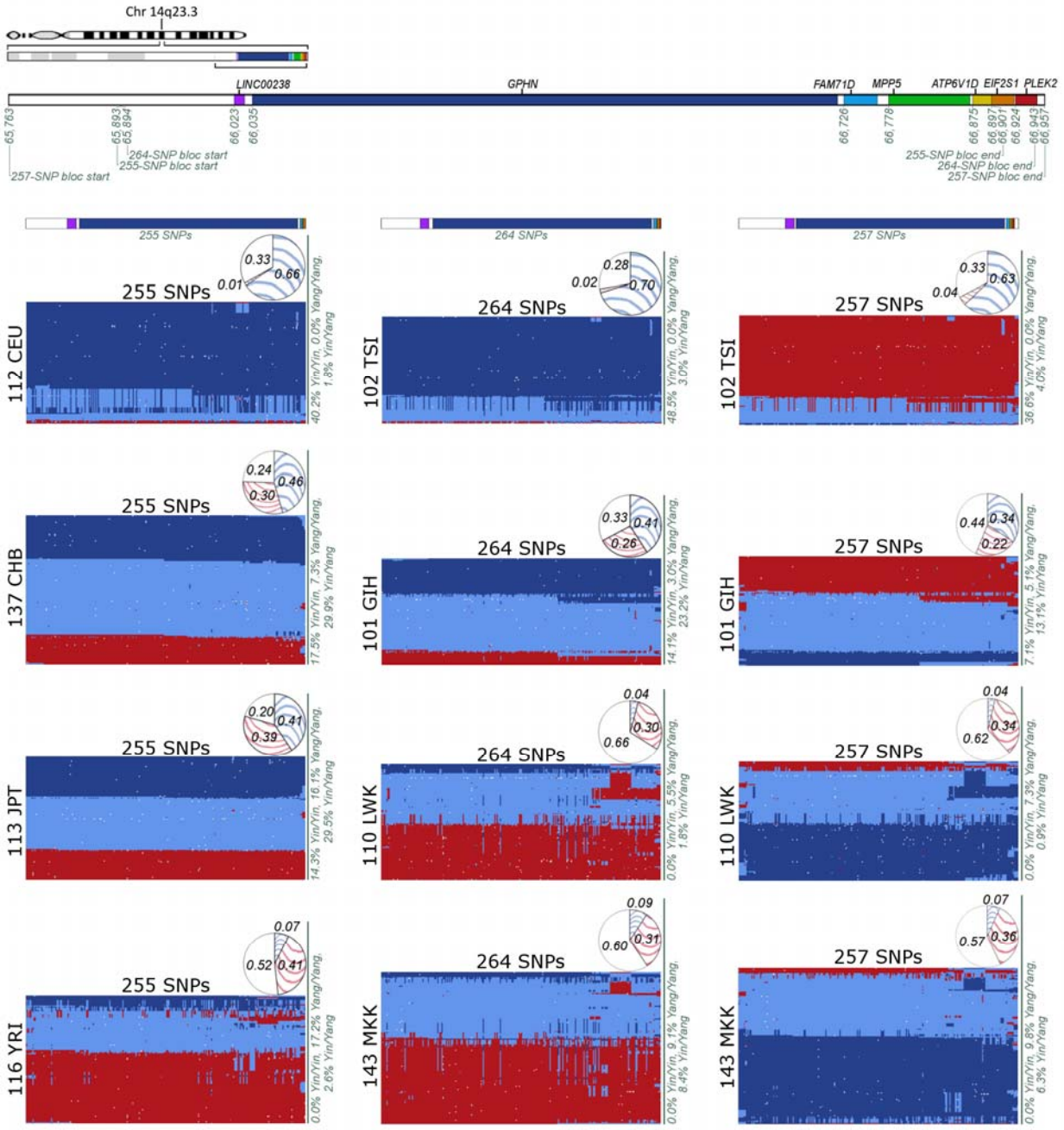


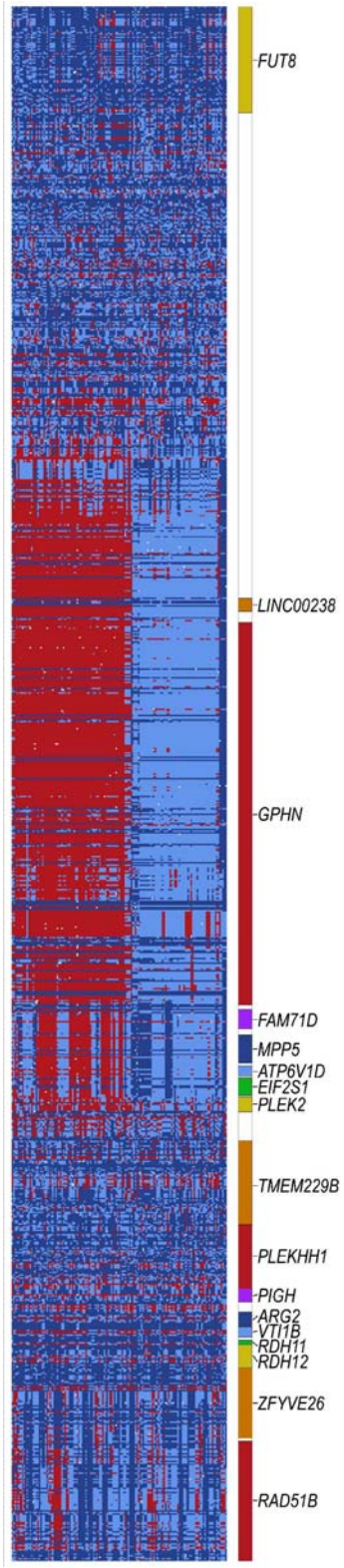
Supplementary Figure 1 | Second bloc identified in second network. BlocBuster identified two blocs during the analysis of the GIH, LWK, MKK, and TSI populations. This bloc was essentially a mirror image of the first bloc in the network, bearing alternate alleles. Homozygotes for alleles in the bloc are shaded dark blue, as in Fig. 1. However, the bloc alleles here represent the *yang* haplotype, which is the opposite of the other figure. Heterozygotes and homozygotes of the alternate allele are shaded light blue and red, respectively. The pie charts use the original colors for yin and yang haplotype frequencies, i.e. blue for yin haplotypes and red for yang haplotypes. See caption of Fig. 1 for additional details.



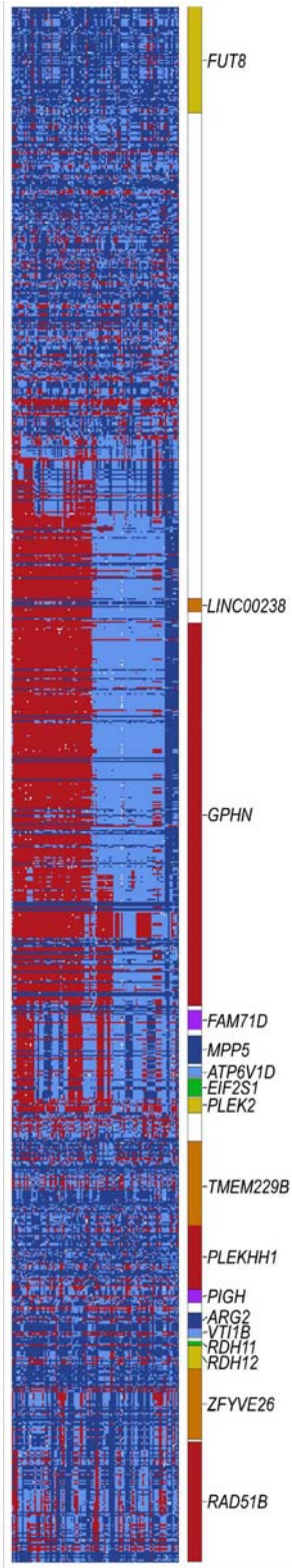
Supplementary Figure 2 | The 1000 Genomes Project bloc. BlocBuster analysis of the yin-yang region of the 1000 Genomes Project data produced a 303-node bloc. Shown are the genotypes for 2,504 individuals from 26 global populations, for which 26.4% and 13.3% are homozygous for the yin and yang haplotypes, respectively, and 20.2% are heterozygotes. See caption of Fig. 1 for additional details.



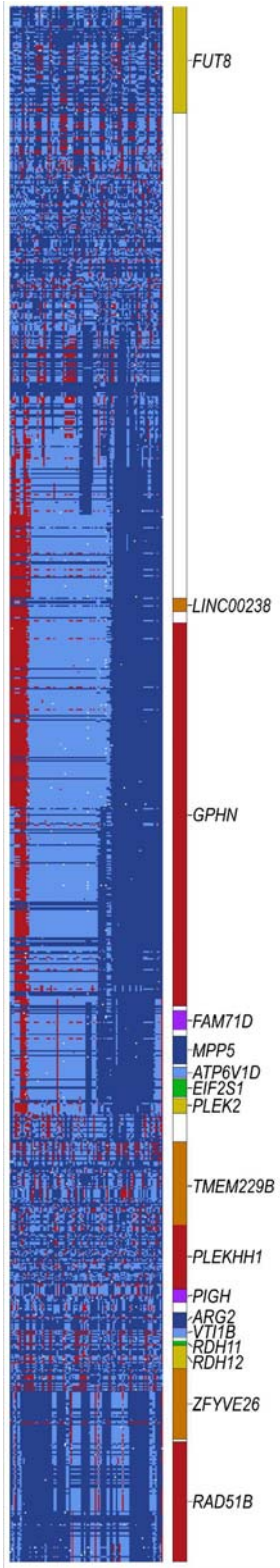
Supplementary Figure 3 | High resolution copy of Figure 1.



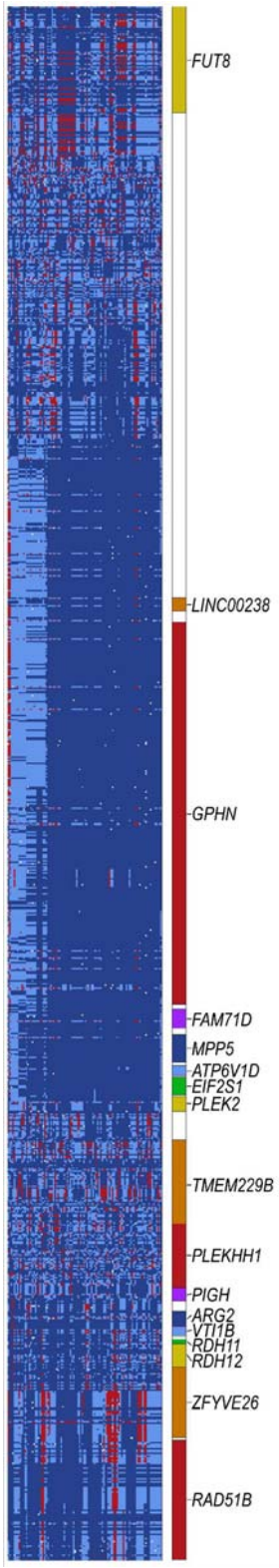
Supplementary Figure 4 | High resolution copy of all SNPs for the MKK population.



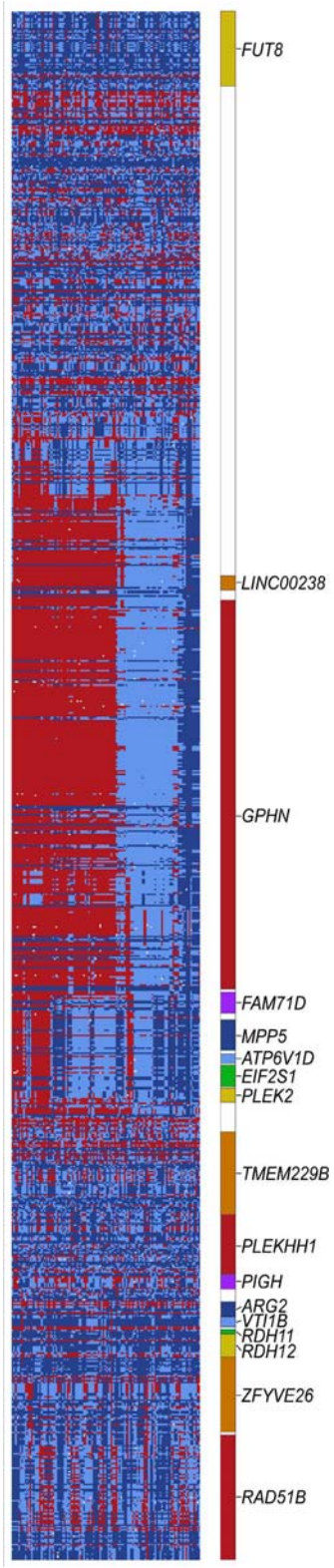
Supplementary Figure 5 | High resolution copy of all SNPs for the LWK population.



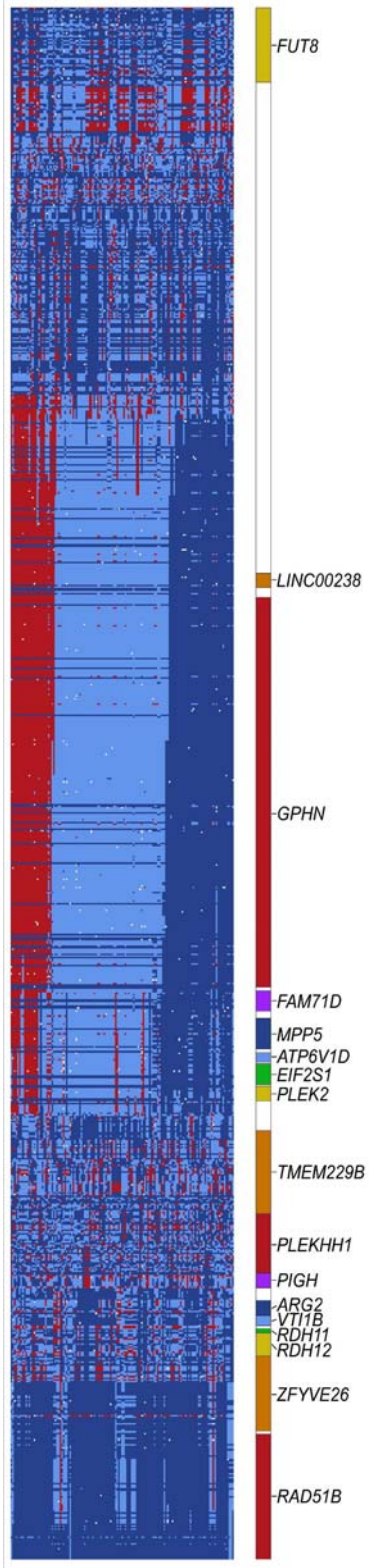
Supplementary Figure 6 | High resolution copy of all SNPs for the GIH population.



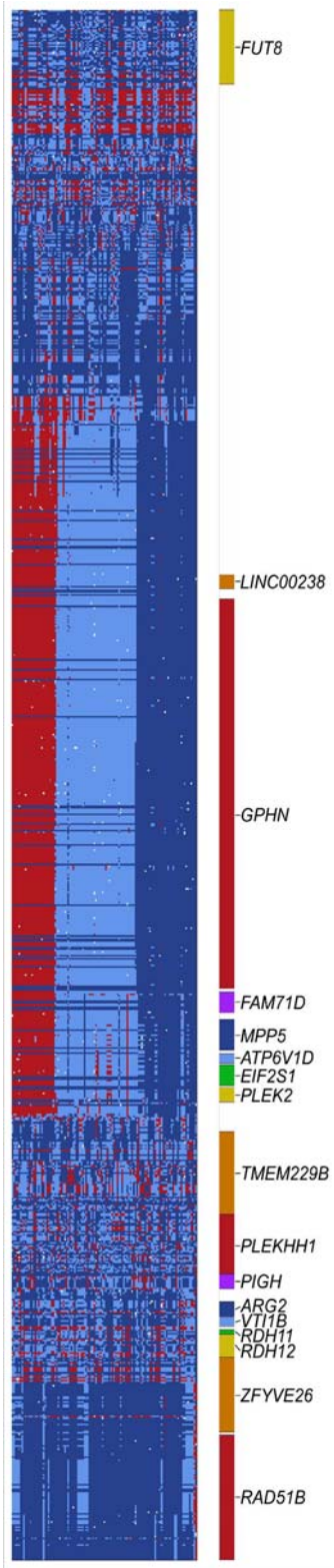
Supplementary Figure 7 | High resolution copy of all SNPs for the TSI population.



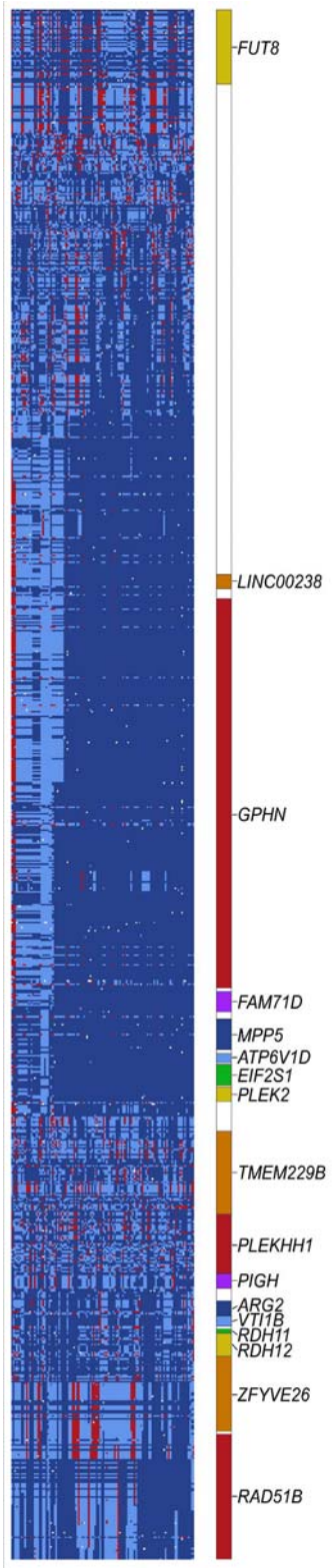
Supplementary Figure 8 | High resolution copy of all SNPs for the YRI population.



Supplementary Figure 9 | High resolution copy of all SNPs for the CHB population.



Supplementary Figure 10 | High resolution copy of all SNPs for the JPT population.



Supplementary Figure 11 | High resolution copy of all SNPs for the CEU population.

Supplementary Note 1

Network details and additional yang haplotype

BlocBuster was used to construct networks comprised of nodes, representing SNP alleles, and edges, representing high correlations between SNP alleles, as described in Methods. The data for four populations were combined and analyzed for each autosomal chromosome. For each network, the number of edges was set equal to the number of SNPs, resulting with an average node degree of one. In both sets of analyses, the CCC threshold to create these networks was substantially higher for chromosome 14 than for the other autosomal chromosomes. Specifically, the first analysis (CEU, CHB, JPT, and YRI) had CCC ≥ 0.7942 and the second (GIH, LWK, MKK, and TSI) had CCC ≥ 0.7796, while the average CCC threshold for the other autosomal chromosomes over both studies was 0.7581.

Chromosome 14 had higher thresholds as there were many edges with exceptionally high CCC scores within the yin-yang region. For the first analysis, 26,762 of the 36,542 edges in the entire network were in the bloc corresponding to the yin haplotype. For the second analysis, there were 40,820 network edges and 30,407 of these resided within the two blocs corresponding to the yin and yang haplotypes.

The largest bloc in the first network, with 255 nodes, corresponded to the yin haplotype. Visual inspection of the genotype values in Fig. 1 of the manuscript indicates a strong yang pattern also. However, the second largest bloc in the network had 84 nodes and was not located in this region. On the other hand, the second network (GIH, LWK, MKK, and TSI) had two large blocs with 264 and 257 nodes, representing the yin and yang haplotypes, respectively.

Note that the CCC threshold for chromosome 14 was substantially higher for the first network than the second. We constructed another network for the CEU, CHB, JPT, and YRI chromosome 14 data with 55,000 edges, resulting with a CCC threshold of 0.7779. The CCC threshold for this new network was similar to the threshold for the original analysis of the GIH, LWK, MKK, and TSI chromosome 14 data. The new network included a bloc of 159 nodes spanning the yin-yang region (position 65.9 Mb to 66.7 Mb) and the SNP alleles corresponded to the yang haplotype. This bloc had 15 nodes in the original network and grew to 159 nodes when the CCC threshold was relaxed to 0.7779.

These results indicate that the SNP alleles for the yin haplotype were exceptionally inter-correlated for the first analysis, so much so that the corresponding network bloc absorbed 73% of the edges in the entire network. Indeed, only 15 of the SNP alleles corresponding to the yang haplotype were captured in the first network due to the extremely high CCC threshold that was driven by these high inter-correlations.

Supplementary Note 2

Results from 1000 Genomes Project data analysis

The 1000 Genomes Project[1] data were analyzed using BlocBuster (see Methods). The goal of the 1000 Genomes Project is to identify most human genetic variants that appear with at least 1% frequency. In comparison with HapMap, this project includes almost twice as many individuals and many additional populations are represented. However, there is some overlap in the samples between the two projects. In order to maximize the number of individuals genotyped, the 1000 Genomes Project used "light" sequencing (4X) across all individuals and intrinsically imputed missing values while calling genotypes. Due to this low-coverage sequencing, there is a high probability that for each individual only one chromosome was actually sampled at a specific site.

This is important in this context as imputation of missing data prior to computing correlations can generate false positive signals. Such imputations are based upon the assumption of linkage disequilibrium (LD) between missing and identified markers[2]. Imputation can be useful for association studies in which each SNP is considered individually. However, errors introduced are biased toward inflated LD, and LD is a property captured by correlation measures, including CCC. The 1000 Genomes Project is unable to provide information regarding which genotypes were imputed, so in general it is not suitable for BlocBuster or other methods designed to identify correlations. We present the results here just to test the robustness of the HapMap results.

The 1000 Genomes Project data included 13,564 biallelic SNPs in the yin-yang region for 2,504 individuals. A BlocBuster network was constructed representing the 13,564 highest CCC values, producing a bloc with 303 nodes, as shown in Supplementary Figure 2. This bloc has only about 7% more SNPs than the 284 identified by the HapMap analysis, despite the high density of SNP data. This observation is most likely due to the HapMap ascertainment capturing most of the common variants in the region, while the 1000 Genomes Project captured a large quantity of less common variants. (We observed that 88.5% of the 13,564 SNPs in the 1000 Genomes data have minor allele frequencies less than 0.100.)

As shown in Supplementary Figure 2, the yin-yang pattern is strong across the 2,504 individuals, thereby supporting the HapMap results.

Supplementary References

1. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).
2. Halperin, E. & Stephan, D. A. SNP imputation in association studies. *Nature biotechnology* **27,** 349–51 (2009).