Help for selecting population features (GO category)

HGPGD database v1.0

The genetic differences of GO category were reflected by combining the differences of all genes in that category.

1. The population genetic differences of single gene

For each autosome gene region among 11 HapMap populations, we measured the differences of SNPs in each gene region using 11 indicators: (1) allele frequency (2) Fst (3) r^2 (4) Dprime (5) Block number (6) Block size (7) SNP density (8) Haplotype diversity (9) tagSNP percent (10) captured percent (11) average max r^2.

These indicators are mainly related to three main aspects: the allele frequency (allele frequency and Fst), LD pattern (r^2, Dprime, Block number, Block size, SNP density and Haplotype diversity) and transferability of tag SNPs (tagSNP percent, captured percent and average max r^2), that were usually used for comparing samples from different populations [1-6] and reflecting some population genetic characteristics.

Allele frequency differences

We measured the average differences of allele frequency for each gene region between pair-wise HapMap populations. The minor allele in HapMap ASW population was used as the reference. For each gene region, we defined the difference of allele frequency $diff_{maf}(i,j)$ as follows:

$$diff_{maf}(i,j) = \frac{1}{N}\sum_{k=1}^{N} | maf_{k,i} - maf_{k,j} |$$

Where i, j are HapMap populations (1:ASW, 2:CEU, 3:CHB, 4:CHD, 5:GIH, 6:JPT, 7: LWK, 8:CEX, 9:MKK, 10:TSI, 11:YRI). N is the number of SNPs in a gene region. $maf_{k,i}$ is the frequency of the k th SNP in population i, $maf_{k,j}$ is the frequency of the k th SNP in population j. A larger $diff_{maf}$ indicates a higher difference of allele frequency in the gene region among 11 HapMap populations, on the contrary a smaller $diff_{maf}$ indicates a lower difference.

Fst. We measured the average Fst for each gene region between pair-wise HapMap populations. The $diff_{Fst}(i,j)$ were calculated in the same way as $diff_{maf}(i,j)$.

LD pattern differences differences

For each gene region, six indicators about LD pattern were calculated.

r^2 differences (LD coefficient r^2 differences) We calculated pairwise LD coefficient r^2 between all pairwise SNPs (less than 500kb). The differences of r^2 $diff_{r^2}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

Dprime differences (D' differences). We calculated pairwise D' between all pairwise SNPs (less than 500kb). The differences of D' $diff_{Dprime}(i, j)$ between pairwise populations were calculated in the same way as $diff_{maf}(i, j)$.

Block number differences For each gene region, Four Gamete Tests (FGT)[7] was used to identify the haplotype block structure, and the block number within the gene region was calculated. The differences of block number $diff_{block\_num}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

Block size differences The average size of blocks within the gene region was calculated. The differences of average block size $diff_{block\_size}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

SNP density differences The average SNP density of blocks within the gene region was calculated. The differences of average SNP density of blocks $diff_{SNP\_dens}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

Haplotype diversity differences For each block in each gene region, haplotype diversity[4] was computed as $h = (1 - \sum x_i^2)n/(n-1)$, where x_i was the frequency of a given haplotype and n was the number of samples, and average haplotype diversity was defined as the average value of haplotype diversity in block regions. The differences of average haplotype diversity $diff_{hap\_div}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

In this study, haploview v4.1[8] was used to identify haplotype block and to estimate haplotype frequency by expectation-Maximization (EM) algorithm.

Transferability of tagSNP differences

There were three indicators about the transferability of tagSNP.

TagSNP percent differences For each gene region, an aggressive tagging strategy by TAGGER panel in haploview was used to identify tagSNPs (r^2 threshold is 0.8). The tag percent was defined as the number of tagSNPs divided by the total number of SNPs in a gene region. The differences of tagSNP percent $diff_{\text{tag\_perc}}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

Captured percent differences For example, for ASW population, if an ASW SNP exhibited pairwise $r^2>0.8$ with at least one tagSNP selected from the CEU population, then the SNP was defined as captured SNP by CEU panel in the ASW population[4], and captured percent was defined as the number of captured SNPs divided by the total number of SNPs in ASW population. The differences of captured percent $diff_{\text{Cap\_perc}}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

Average maximum r^2 differences For each gene region, average maximum r^2 was defined as the average value of the maximum r^2 between tagSNPs in one HapMap population and SNPs captured by these tagSNPs in another population. Captured percent and Average maximum r^2 were used to evaluate the efficiency of tagSNPs in one population to capture SNPs in another population. The differences of average maximum r^2 $diff_{\max\_r^2}(i, j)$ were calculated in the same way as $diff_{maf}(i, j)$.

2. The population genetic differences of GO categories

The genetic differences of GO category were reflected by combining the differences of all genes in that GO category. Some previous studies have shown that genes assigned the same functional set are more likely to share certain biological characteristics than random sets of genes [9, 10].

For each GO category, we assigned the same weight to genes in the GO category and calculated genetic difference scores for each 10 indicator separately.

For example, for each GO category, we calculated the allele frequency difference as follows:

$$D_{maf}(i, j) = \sum_{k=1}^{M} \frac{1}{M} diff_{maf}(i, j),$$

where i, j are HapMap populations (1:ASW, 2:CEU, 3:CHB, 4:CHD, 5:GIH, 6:JPT, 7:LWK, 8:MEX, 9:MKK, 10:TSI, 11:YRI), M is the gene number in the GO category. $D_{maf}(i, j)$ was used to measure the allele frequency difference between population i and population j.

The differences of other 9 indicators were calculated in the same way as $D_{maf}(i, j)$.

1. De Bakker, P.I., R.R. Graham, D. Altshuler, et al., *Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations.* Pac Symp Biocomput, 2006: p. 478-86.

2. Service, S., J. DeYoung, M. Karayiorgou, et al., *Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies.* Nat Genet, 2006. **38**(5): p. 556-60.
3. Marvelle, A.F., L.A. Lange, L. Qin, et al., *Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples.* J Hum Genet, 2007. **52**(9): p. 729-37.
4. Ribas, G., A. Gonzalez-Neira, A. Salas, et al., *Evaluating HapMap SNP data transferability in a large-scale genotyping project involving 175 cancer-associated genes.* Hum Genet, 2006. **118**(6): p. 669-79.
5. Xing, J., D.J. Witherspoon, W.S. Watkins, et al., *HapMap tagSNP transferability in multiple populations: general guidelines.* Genomics, 2008. **92**(1): p. 41-51.
6. Lundmark, P.E., U. Liljedahl, D.I. Boomsma, et al., *Evaluation of HapMap data in six populations of European descent.* Eur J Hum Genet, 2008. **16**(9): p. 1142-50.
7. Wang, N., J.M. Akey, K. Zhang, et al., *Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation.* Am J Hum Genet, 2002. **71**(5): p. 1227-34.
8. Barrett, J.C., B. Fry, J. Maller, et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.
9. Aerts, S., D. Lambrechts, S. Maity, et al., *Gene prioritization through genomic data fusion.* Nat Biotechnol, 2006. **24**(5): p. 537-44.
10. Holmans, P., E.K. Green, J.S. Pahwa, et al., *Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder.* Am J Hum Genet, 2009. **85**(1): p. 13-24.