BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 20 2005, pages 3938–3939
doi:10.1093/bioinformatics/bti649

Genetics and population analysis

HAPLOT: a graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations

Sheng Gu*, Andrew J. Pakstis and Kenneth K. Kidd

Department of Genetics, Yale University School of Medicine, New Haven, USA

Received on July 13, 2005; revised on August 24, 2005; accepted on August 25, 2005
Advance Access publication August 30, 2005

ABSTRACT

Summary: Understanding of human variation relevant to association studies can benefit from population comparison, especially comparing populations in the same geographical region. Variations in linkage disequilibrium patterns, in tagSNP sets, and in SNP heterozygosities among populations can be used to infer the evolutionary pattern. We present here a win32 system based Perl/Tk application for visual comparisons of these variations in different populations.

Availability: The application package is available at http://info.med.yale.edu/genetics/kkidd/programs.html

Contact: sheng.gu@yale.edu

Association studies of single nucleotide polymorphisms (SNPs) impart invaluable knowledge about human evolution and disease-related gene identification. Closely located sites with strong linkage disequilibrium (LD) may define only a few common haplotypes that account for a large proportion of chromosomes. A subset of SNPs, called tagSNPs, that captures the most variation among the haplotypes can be chosen through appropriate statistics.

However, whatever partition algorithm has been applied, the regions of strong LD may not generalize among different populations. It has been reported recently that haplotype block is an oversimplified view of linkage disequilibrium in genomic regions and block structures show significant variation among populations (Liu *et al*., 2004). Thus, 'blocks' may not reflect any fundamental structure of the genome but can be considered regions of stronger LD specific to a population because of its unique demographic history.

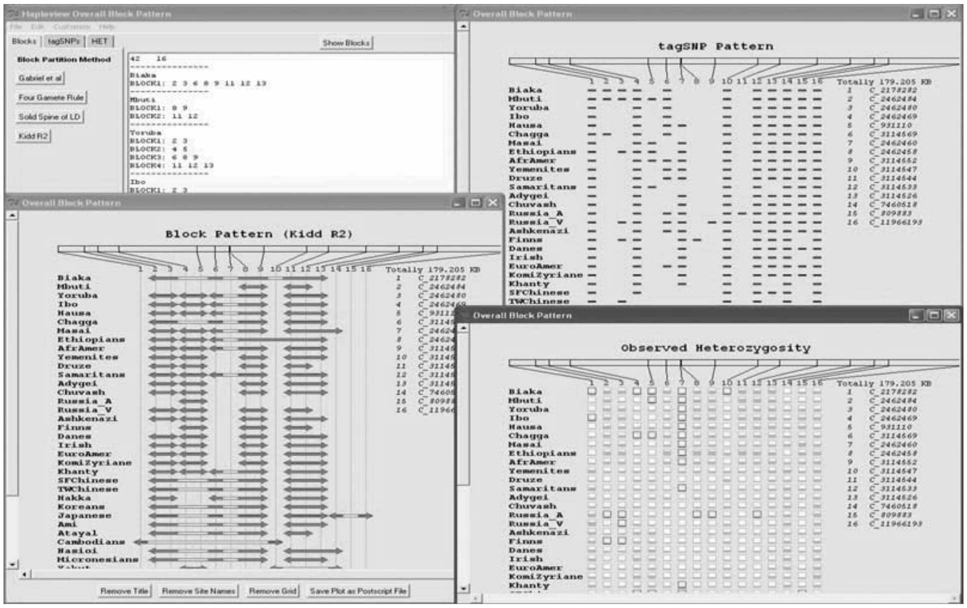
tagSNP selection, which is one focus of the HapMap Project (The International HapMap Consortium, 2003), is most valuable when it carries good transferability (generalizability) among populations, especially populations in the same geographical regions. Therefore, comparative studies of many population samples within and across geographical regions will increase the accuracy of tagSNP selection.

Here, we present a win32 system based application, HAPLOT, for graphical comparisons of haplotype blocks, selected tagSNPs and SNP heterozygosity among populations. HAPLOT is developed in Perl/Tk, with a java application Haploview (Barrett *et al*., 2005) and a C++ program HapBlock (Zhang *et al*., 2005) incorporated for calculations in different modules. Three different input formats of raw data are accepted: a self-defined block file simply constructed by population names and block components, which allows users to incorporate results from any block partition algorithm not included in this application; one or more pedigree files, whose format is compatible with Haploview input (Barrett *et al*., 2005); or a master data file, which is specific to PhenoDB database (Cheung *et al*., 1996) users. An accessory information file, which contains the site names and physical locations, must be provided together with the raw data. An optional file, 'pop.list', if provided, will group populations into user-defined order.

Four different block partition algorithms can be selected to generate block structures. The block definitions based on the LD measure D′ confidence interval (Gabriel *et al*., 2002), the four-gamete test (Wang *et al*., 2002), and the solid spine of LD (Barrett *et al*., 2005) are embedded in Haploview and are called by HAPLOT directly for block pattern plotting, while a new block definition from the LD measure r^2 is an internally developed method that initiates and extends a block according to the pairwise and grouped r^2 values. The algorithm starts a block by selecting the pair of adjacent SNPs with the highest r^2 value (no less than α) and extends that block if the average r^2 value between an adjacent site and current block members is above β and all the individual r^2 values are above γ. Here, $\alpha > \beta > \gamma$, and by default, α, β, γ are set as 0.4, 0.3 and 0.1, respectively. After the first block is identified, a new pair of adjacent SNPs with the next highest r^2 value (no less than α) is used to start a new block accretion process. Occasionally, an SNP can be assigned to either one of two adjacent blocks. The ambiguity is resolved by giving the earlier identified block (the starting pair of SNPs has higher r^2 values) the priority to possess that target SNP. The α, β, γ values given were selected because they show better consistency of block patterns among populations within the same geographical region in preliminary studies. The graphical representation of a block uses a solid red line with double arrows for delimitation. If inside a block an SNP is below the minimum heterozygosity threshold or fails the Hardy–Weinberg test, the corresponding part of the segment is not coloured.

Four tagSNP selection methods are available. The haplotype diversity method (Clayton, 2001 http://www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf) and the haplotype entropy method (Nothnagel *et al*., 2003) are implemented in Hap Block, and the pairwise and aggressive LD methods are implemented in Haploview. Thus, these algorithms are loaded directly by HAPLOT for tagSNP selection. However, ambiguous tagSNP sets (any one of several SNPs in a specific haplotype can be used to

*To whom correspondence should be addressed.



```
42   16
---------------
Biaka
BLOCK1: 2 3 6 8 9 11 12 13
---------------
Mbuti
BLOCK1: 8 9
BLOCK2: 11 12
---------------
Yoruba
BLOCK1: 2 3
BLOCK2: 4 5
BLOCK3: 6 8 9
BLOCK4: 11 12 13
---------------
Ibo
BLOCK1: 2 3
```

Fig. 1. A graphical presentation of haplotype blocks, selected tagSNPs and SNP heterozygosity in different populations.

represent that haplotype variation) could occur in all types of algorithms. Since HapBlock provides five quantities (the percentage of uniquely distinguished haplotypes, the proportion of haplotype diversity, the percentage of haplotype entropy, the minimum value of haplotype prediction strength for common haplotypes and the minimax value of pairwise LD-measure r^2) with each tagSNP set, HAPLOT resolves ambiguity by choosing the tagSNP set with the largest sum of these five equal-weighted quantities after standardizing. For LD-based methods, HAPLOT resolves ambiguity by grouping ambiguous tagSNP sets from all the populations and then selecting the SNP that occurs most frequently in all ambiguous tagSNP sets as the representative SNP. After a representative SNP is chosen, all related ambiguous sets are discarded, and this representative SNP is added to the unambiguous tagSNP list. Then the next most frequent tagSNP in the remaining ambiguous tagSNP sets is located for another processing cycle until no ambiguity exists in any population. The graphical representation of a tagSNP uses a short blue segment.

Both observed and predicted heterozygosity of all SNPs in all populations can be presented in tabular and graphic formats. SNPs below a self-defined threshold are highlighted in red in the plot, and the value of heterozygosity is reflected by the level of the colour filling in a unit box. In addition, SNPs that fail the Hardy–Weinberg test are pointed out.

Parameters for different algorithms or graphical representation can be user-defined. The tabular result can be exported to a text file and the plot can be saved as a postscript file. The graphical output formats for haplotype block, tagSNP, and SNP heterozygosity patterns are illustrated in a coloured plot (Fig. 1). There patterns are generated from the same source file and thus can be compared in parallel.

ACKNOWLEDGEMENTS

We thank Ronny C. Woodruff for comments. This work was supported in part by NIH GM57672.

Conflict of Interest: none declared.

REFERENCES

Barrett,J.C. *et al*. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Carlson,C.S. *et al*. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *Am. J. Hum. Genet*., **74**, 106–120.

Cheung,K.H. *et al*. (1996) PhenoDB: an integrated client/server database for linkage and population genetics. *Comput. Biomed. Res*., **29**, 327–337.

Clayton,D. (2001) .

Gabriel,S.B. *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Liu,N. *et al*. (2004) Haplotype block structures show significant variation among populations. *Genet. Epid*., **27**, 385–400.

Nothnagel,M. *et al*. (2003) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum. Hered*., **54**, 186–198.

The International HapMap Consortium (2003), The International HapMap Project. *Nature*, **426**, 789–796.

Wang,N. *et al*. (2002) A dynamic programming algorithm for haplotype partitioning. *Proc. Natl. Acad. Sci. USA*, **99**, 7335–7339.

Zhang,K. *et al*. (2005) HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics*, **21**, 131–134.