BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 1 2005, pages 131–134
doi:10.1093/bioinformatics/bth482

HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms

Kui Zhang[1], Zhaohui Qin[2], Ting Chen[3], Jun S. Liu[4], Michael S. Waterman[3] and Fengzhu Sun[3,*]

[1]Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA, [2]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA [3] Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, 1042 W. 36th Place DRB-288, Los Angeles, CA 90089–1113, USA and [4]Department of Statistics, Harvard University, Cambridge, MA 02138, USA

Received on June 1, 2004; revised on July 20, 2004; accepted on August 11, 2004
Advance Access publication August 27, 2004

ABSTRACT

Summary: Recent studies have revealed that linkage disequilibrium (LD) patterns vary across the human genome with some regions of high LD interspersed with regions of low LD. Such LD patterns make it possible to select a set of single nucleotide polymorphism (SNPs; tag SNPs) for genome-wide association studies. We have developed a suite of computer programs to analyze the block-like LD patterns and to select the corresponding tag SNPs. Compared to other programs for haplotype block partitioning and tag SNP selection, our program has several notable features. First, the dynamic programming algorithms implemented are guaranteed to find the block partition with minimum number of tag SNPs for the given criteria of blocks and tag SNPs. Second, both haplotype data and genotype data from unrelated individuals and/or from general pedigrees can be analyzed. Third, several existing measures/criteria for haplotype block partitioning and tag SNP selection have been implemented in the program. Finally, the programs provide flexibility to include specific SNPs (e.g. non-synonymous SNPs) as tag SNPs.

Availability: The HapBlock program and its supplemental documents can be downloaded from the website http://www.cmb.usc.edu/~msms/HapBlock

Contact: fsun@usc.edu

Genome-wide association methods based on linkage disequilibrium (LD) offer a promising approach to detect genetic variation responsible for human common diseases. Single nucleotide polymorphism (SNP) markers are preferred for disease association studies because of their high abundance along the human genome, the low mutation rate and accessibility to high-throughput genotyping. However, the current throughput of technology is inadequate for genotyping all existing SNPs for a large number of samples. Thus, the selection of a maximally informative set of SNPs (tag SNPs) for genome-wide association studies has attracted much attention recently. Several large-scale studies for dissecting LD patterns across the human genome based on SNPs have revealed that the LD patterns vary greatly across the human genome with some regions of high LD interspersed with regions of low LD (Gabriel *et al*., 2002; Patil *et al*., 2001). In those high LD regions, which are referred to as blocks in the literature, only a small number of SNPs are sufficient to capture most of haplotype structure (Johnson *et al*., 2001; Patil *et al*., 2001). Therefore, it is desirable to apply the knowledge of such block-like patterns in selecting a set of tag SNPs, such that the genotyping burden could be reduced without much loss of power for disease association studies.

We have developed a suite of computer programs, named HapBlock, for haplotype block partitioning and tag SNP selection. Compared to other available similar programs, our suite of programs has several distinct features.

First, our program incorporates a set of dynamic programming algorithms (Zhang *et al*., 2002, 2003, 2004a). The program depends on two main criteria: the definition of candidate haplotype blocks and a criterion for finding tag SNPs within blocks. The primary objective of the programs is to minimize the total number of tag SNPs across the human genome. The secondary objective is to minimize the total number of blocks among those block partitions with the minimum number of tag SNPs. Zhang *et al*. (2002) developed a dynamic programming algorithm to search all possible block partitions for haplotype data. This algorithm guarantees an optimal solution. Zhang *et al*. (2003) also proposed two dynamic

*To whom correspondence should be addressed.

programming algorithms for tag SNP selection with limited resources. With limited resources, we want to find a block partition with at most a given number of tag SNPs to cover as much of the genome as possible. This can be formulated as two equivalent dual problems: block partition with a Fixed Genome Coverage using the minimum number of tag SNPs (FGC) and block partition with a Fixed number of Tag SNPs that can cover the maximum length of genome (FTSNP). A parametric dynamic programming algorithm and a two-dimensional dynamic programming algorithm are developed to solve the FGC and the FTSNP problem, respectively. For details of these algorithms, please refer to Zhang *et al*. (2003).

Second, haplotypes as well as genotypes from unrelated individuals and general pedigrees can be analyzed in our program. The throughput of current haplotyping technology is insufficient to determine haplotypes from diploid individuals in a large-scale study across the whole genome such as the one reported by Patil *et al*. (2001). Therefore, large-scale genotype data other than haplotype data will be generated in many projects, such as in the HapMap project. When only genotype data are available, a naive approach is to infer haplotypes based on all SNPs and their frequencies and take them as input. However, this inference may not be feasible and reliable due to a large number of SNPs and low LD in some regions. Thus, a different strategy is employed in our program (Zhang *et al*., 2004a). The haplotypes and their frequencies are inferred for each subset of consecutive SNPs that can form a potential block. The haplotypes inferred within potential blocks are reliable due to high LD. When genotypes from unrelated individuals are available, a PL-EM algorithm (Qin *et al*., 2002) is employed. Both the most likely haplotype pairs assigned to individuals and haplotypes and their frequencies can be used. When genotypes from general pedigrees are available, haplotypes and their frequencies are inferred by a set of logic-rules and the PL-EM algorithm (Zhang *et al*., 2004b) under the assumption of no recombination events. The assumption of no recombination seems to be too strict at a first glance. However, it is reasonable for dense SNP maps and enables much more rapid computation.

Third, several definitions for haplotype blocks and tag SNPs are supported by the program. It is worth noting that a single SNP is still considered as a block in this paper. Currently, three different block definitions are implemented:

(1) *The coverage of common haplotypes*. It follows the definition proposed by Patil *et al*. (2001), in which at least a certain percentage of observed or inferred haplotypes must be common haplotypes.

(2) *LD-based blocks*. This definition is based on LD measure D' and is similar to those used in Zhang and Jin (2003) and Gabriel *et al*. (2002). In each block, we require at least a certain proportion of SNP pairs having strong LD (the pair-wise $|D'|$ greater than a threshold).

(3) *No historical recombination*. This definition is proposed by Wang *et al*. (2002) and has been implemented in HaploBlockFinder (Zhang and Jin, 2003). A set of consecutive SNPs is defined as a block if there are no historical recombination events, which is based on the four-gamete test.

A set of variant approaches for tag SNP selection has been implemented in the program. There are five definitions for tag SNPs so far:

(1) *Common haplotypes*. This method was first proposed by Patil *et al*. (2001) and used by Zhang *et al*. (2002). In a block, the minimum set of SNPs that can uniquely distinguish a certain percentage of all the haplotypes is considered as a set of tag SNPs.

(2) *Haplotype diversity* (Johnson *et al*., 2001). Here, the minimum set of SNPs that can account for a certain percentage of overall haplotype diversity is defined as a set of tag SNPs.

(3) *Haplotype entropy*. Entropy has recently been used for haplotype block partitioning (Nothnagel *et al*., 2002). If there are n haplotypes and the frequency of haplotype i is denoted by p_i, then the entropy of these haplotypes is defined as $S = -\sum_{i=1}^{n} p_i \log p_i$. Here, we extend this concept to tag SNP selection: the set of tag SNPs is the minimum set of SNPs that can account for a certain percentage of overall haplotype entropy.

(4) *Haplotype determination coefficient* R_h^2. Stram *et al*. (2003) proposed a formal measure, R_h^2, to characterize the uncertainty in the prediction of haplotypes from genotype data and used it for tag SNP selection. In this method, R_h^2 for each common haplotype is calculated based on a subset of SNPs over all SNPs within a block. The minimum R_h^2 among all common haplotypes is taken as the overall 'haplotype prediction strength'. The minimum set of SNPs with the overall haplotype prediction strength exceeding a pre-specified threshold is taken as a set of tag SNPs.

(5) *LD measure* r^2. The LD measure r^2 has been used for tag SNP selection (Carlson *et al*., 2004; Zhang and Jin, 2003) because the statistical power of association studies is proportional to the value of r^2. Here, for any given subset of SNPs within a block, all pair-wise r^2 values between the SNPs in this subset and the SNPs absent in this subset are calculated. For a given SNP absent in the subset, we take the maximum value of r^2 as its individual prediction power. The minimum value over all SNPs absent in this subset is taken as the overall prediction power. The minimum set of SNPs with prediction power exceeding a pre-specified threshold is considered as a set of tag SNPs.

The dynamic programming algorithm will search all possible block partitions. If genotype data are used, we use the PL-EM algorithm to infer the haplotypes of individuals and to estimate the haplotype frequencies in a sample within each potential block. In addition, we use the enumeration method to search for tag SNPs within each block. Therefore, our program runs slower than greedy algorithms but guarantees that we find a block partition with the minimum number of tag SNPs. To test the program, we use a dataset generated by Daly *et al*. (2001). This dataset contains 103 SNPs with minor allele frequency $>5\%$ based on genotypes of 129 triads. A detailed comparison using our program based on the inferred transmitted haplotypes and the genotypes of offspring can be found in Zhang *et al*. (2004a). Here, we focus on the speed of the program. We perform our analysis based on three types of data: the transmitted haplotypes inferred from trios, the genotypes of offspring and the genotypes of all trios. The common haplotypes are those having frequency $>5\%$. We define a consecutive set of SNPs as a block if the common haplotypes account for at least 80% of all haplotypes and the tag SNPs are the minimum set of SNPs that can distinguish at least 80% of all haplotypes. On a Dell Workstation Precision 360 with a 2.60 GHz Pentium 4 CPU and 1024 MB RAM, the three datasets described above take about 36, 1566, and 1080 s to find the optimal solution, respectively.

The programs have several additional features. First, all optimal sets of tag SNPs are listed with five summary quantities based on uniquely distinguished common haplotypes, haplotype diversity, haplotype entropy, haplotype determination coefficient R_h^2, and LD measure r^2. This would be very helpful in identifying a most informative set of tag SNPs for specific applications. Second, the program provides an option to include some specific SNPs, such as non-synonymous SNPs and other important functional SNPs from previous studies, into the set of tag SNPs. This feature has also been implemented in SNPtagger (Ke and Cardon, 2003). Third, the haplotypes and their frequencies within each block are listed and can be used by other programs, such as SNPtagger (Ke and Cardon, 2003) for further inspection.

Our interest in haplotype block analysis is for its potential utility in association studies where it can reduce the genotyping burden while preserving high power for association studies. The program provides a platform to detect haplotype blocks and select tag SNPs using a suite of dynamic programming algorithms with various criteria, which would greatly facilitate such analysis. New methods for haplotype block partitioning and tag SNP selection will be brought out in the future and we would like to implement them in this program. At the same time, few studies have been carried out to assess the power of association studies using tag SNPs chosen based on different methods. Our program offers a unified platform to conduct such studies.

ACKNOWLEDGEMENT

We thank two anonymous reviewers for their thoughtful and constructive comments. This work is partially supported by NSF DMS-0104129, NIH R01-HG02518 (ZQ, JSL) and NIH P50 HG 002790, RR16522, NSF EIA-0112934, the University of Southern California (K.Z., M.S.W., T.C. and F.S.), NIH R01ES09912 and NIH U54CA100949 (K.Z.).

REFERENCES

Carlson,C.S., Eberle,M.A., Rieder,M.J., Yi,Q., Kruglyak,L. and Nickerson,D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analysis using linkage disequilibrium. *Am. J. Hum. Genet*., **74**, 106–120.

Daly,M.J., Rioux,J.D., Schaffner,J.F., Hudosn,T.J. and Lander,E.S. (2001) High-resolution haplotype structure in the human genome. *Nat. Genet*., **29**, 229–232.

Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M. *et al*. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.

Johnson,G.C.L., Esposito,L., Barratt,B.J., Smith,A.N., Heward,J., Di Genova,G., Ueda,H., Cordell,H.J., Eaves,I.A., Dudbridge,F. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet*., **29**, 233–237.

Ke,X. and Cardon,L.R. (2003) Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, **19**, 287–288.

Nothnagel,M., Furst,R. and Rohde,K. (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum. Hered*., **54**, 186–198.

Patil,N., Berno,A.J., Hinds,D.A,, Barrett,W.A., Doshi,J.M., Hacker,C.R., Kautzer,C.R., Lee,D.H., Marjoribanks,C., McDonough,D.P. *et al*. (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, **294**, 1719–1723.

Qin,Z., Niu,T. and Liu,J. (2002) Partitioning-Ligation-Expectation-Maximization Algorithm for haplotype inference with single-nucleotide polymorphisms. *Am. J. Hum. Genet*., **71**, 1242–1247.

Stram,D.O., Haiman,C.A., Hirschhorn,J.N., Altshuler,D., Kolonel,L.N., Henderson,B.E. and Pike,M.C. (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using preliminary sample of unrelated subjects with an example from the multiethic cohort study. *Hum. Hered*., **55**, 27–36.

Wang,N., Akey,J.M., Zhang,K., Chakraborty,K. and Jin,L. (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet*., **71**, 1227–1234.

Zhang,K., Deng,M., Chen,T., Waterman,M.S. and Sun,F. (2002) A dynamic programming algorithm for haplotype partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.

Zhang,K. and Jin,L. (2003) HaploBlockFinder: haplotype block analysis. *Bioinformatics*, **19**, 1300–1301.

Zhang,K., Qin,Z., Ting,C., Waterman,M.S., Liu,J.S. and Sun,F. (2004a) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res*., **14**, 908–916.

Zhang,K., Sun,F., Waterman,M.S. and Chen,T. (2003) Dynamic programming algorithms for haplotype block partitioning: applications to human chromosome 21 haplotype data. *Am. J. Hum. Genet*., **73**, 63–73.

Zhang,K., Sun,F. and Zhao,H. (2004b) HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics*, (in press).