BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 1 2007, pages 129–131
doi:10.1093/bioinformatics/btl532

Genetics and population analysis

Combining functional and linkage disequilibrium information in the selection of tag SNPs

P. C. Sham[1,2,3,*], S. I. Ao[4], J. S. H. Kwan[1,2], P. Kao[2], F. Cheung[2], P. Y. Fong[2] and M. K. Ng[5]

[1]Department of Psychiatry, [2]Genome Research Centre, [3]SGDP Centre, Institute of Psychiatry, King's College, London, UK, [4]Department of Mathematics, University of Hong Kong, Shatin, Hong Kong and [5]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Received on May 7, 2006; revised on October 3, 2006; accepted on October 12, 2006
Advance Access publication October 23, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: We have developed an online program, WCLUSTAG, for tag SNP selection that allows the user to specify variable tagging thresholds for different SNPs. Tag SNPs are selected such that a SNP with user-specified tagging threshold C will have a minimum R2 of C with at least one tag SNP. This flexible feature is useful for researchers who wish to prioritize genomic regions or SNPs in an association study.

Availability: The online WCLUSTAG program is available at http://bioinfo.hku.hk/wclustag/

Contact: mng@math.hkbu.edu.hk

1 INTRODUCTION

There are two main approaches to selecting genetic markers in association studies of complex diseases. The first is a direct or functional approach, in which polymorphisms are selected if they cause a change in the amino acid sequence or expression of candidate genes. The second is an indirect or positional approach in which markers in a particular region or the whole genome are systematically screened, based on that they may be in linkage disequilibrium (LD) with disease-related functional variants. For the second approach, efficiency can be improved by recognizing the redundancy between nearby markers through the presence of LD. A subset of SNPs, called tag SNPs, can be selected for genotyping and analysis with minimal loss of information (Halldorsson *et al*., 2004; Johnson *et al*., 2001). Several programs for tag SNP selection are now available, including Tagger (de Bakker *et al*., 2005), HapBlock (Zhang *et al*., 2005) and CLUSTAG (Ao *et al*., 2005).

In this report, we propose novel tag SNP selection algorithms (implemented in the program WCLUSTAG) that take account of functional as well as LD information. More importance is attached to some SNPs than others, based on their positions within coding, regulatory regions or splice sites. We also describe methods to address other practical issues: some SNPs may be more readily assayed than others under the proposed genotyping platform, and some SNPs may have been genotyped in the sample.

WCLUSTAG is developed from the program CLUSTAG by adding the variable tagging threshold and other facilities, and a user-friendly interface. The original method in CLUSTAG was based on agglomerative hierarchical clustering, which starts from a square matrix of pairwise distances between the objects to be clustered. The two clusters with the smallest inter-cluster distance are successively merged until all the objects have been merged into a single cluster. For two SNPs, an appropriate distance measure for LD tagging is $1 - R^2$, where R^2 is the squared correlation between the SNPs. As various forms of agglomerative clustering differ in their definitions of the distance between the two clusters (each of which may contain more than one object), we previously proposed our definition for inter-cluster distance as follows:

(1) For each SNP belonging to either cluster, find the maximum distance (i.e. $1 - R^2$) from it to all the other SNPs in the two clusters.

(2) The smallest of these maximum distances is defined as the distance between the two clusters.

(3) The corresponding SNP is defined as the tag SNP of the newly merged cluster.

In this method, called minimax clustering, setting a cutoff merging distance of C for terminating the algorithm would ensure that no SNP is further than C away from the tag SNP in its cluster. In addition, two other tag SNP selection procedures were implemented in CLUSTAG, a complete linkage clustering method (Byng *et al*., 2003) and a set-cover algorithm similar to the greedy algorithm (Carlson *et al*., 2004). We showed that complete linkage clustering results in a greater number of clusters, while the set-cover method is similar to minimax clustering in terms of the number of tag SNPs but produces less compact clusters (as measured by the average of the distances, $1 - R^2$ between all SNPs and their assigned tag SNPs).

The modification in WCLUSTAG allows the tagging threshold, C, as specified by the user, to be variable among SNPs. Factors that might influence the tagging threshold include positional and functional considerations, as well as other practical issues, such as assay quality and whether the SNP has been genotyped. For instance, C might be set at a high value (e.g. 0.8) for SNPs within the coding or

[*] To whom correspondence should be addressed

regulatory regions of genes expressed in a certain tissue, while a low value (e.g. 0.4) is given to the remaining SNPs.

One complication of this modification is the asymmetry between two SNPs with different values of C e.g. if a coding SNP is given a C of 0.8, and another non-coding SNP is given a C of 0.4, and the R^2 between these two SNPs is 0.6, then it is clear that the first SNP can serve as tag SNP for the second, but not the other way round. Fortunately, the clustering program is able to handle an asymmetric distance matrix, in which the distance from object i to object j is not necessarily the same as the distance from object j to object i. Because of this, the desired extension can be achieved by the following modifications to our clustering algorithm:

(1) A user-defined value of C is provided for each marker.

(2) The distance from marker i to marker j is defined as $C_j - R_{ij}^2$, where C_j is the value of C specified for marker j. If $C_j - R_{ij}^2 < 0$, then marker i can serve as a tag SNP for marker j.

(3) This asymmetric distance matrix is subjected to the minimax clustering method with the cutoff merging distance set at zero. In order words, a cluster is formed if there is a tag SNP, which has a distance 0 or less with each cluster member.

Other modifications to our algorithm are to include SNPs that have been genotyped as well as exclude those that cannot be assayed, and these are done by changing certain elements in the matrix of similarities [R_{ij}^2]. Thus, if marker t has been already genotyped, then all elements of column t in the matrix are set zero, except for the diagonal element, which remains one. This ensures marker t is not tagged by other markers except its own and therefore must be included as one of the tag SNPs in our algorithm. Likewise, if marker t is problematic for assay design, then all elements of row t in the matrix are set zero and hence marker t can never serve as one of the tag SNPs. However, these settings alone do not always ensure the tagging of all SNPs that cannot be assayed; to do this it may be necessary to force the selection of certain SNPs (those required for tagging non-assayable SNPs; see the WCLUSTAG website for details).

Similar modifications can be applied to the set-cover algorithm—marker i can serve as tag SNP for marker j if the condition $C_j - R_{ij}^2 < 0$ is fulfilled. The algorithm would initially select all SNPs that have been already genotyped, and remove the markers tagged by these SNPs. Then the greedy algorithm proceeds as usual, except the exclusion of SNPs that have problems with assay design from the set of possible tag SNPs. As with the clustering algorithm, it is necessary to ensure that tag SNPs for 'non-assayable' SNPs are selected.

The new algorithms were applied to the CEPH sample genotype data from the International Haplotype Map Project. The ENCODE regions were selected since data were available for all known SNPs in these regions. Intragenic regions were identified from the start and end points of the coding sequences for the 33 K Ensemble genes in NCBI build 34. SNPs in these intragenic regions (representing approximately one-third of all SNPs) were given a tagging threshold of 0.8, while others were given a threshold of 0.4. Compared to a uniform tagging threshold of 0.8, setting these variable thresholds reduced the number of tag SNPs by 10–60% in the 10 ENCODE regions, depending on the proportion of the SNPs in the region that are intragenic (Fig. 1).

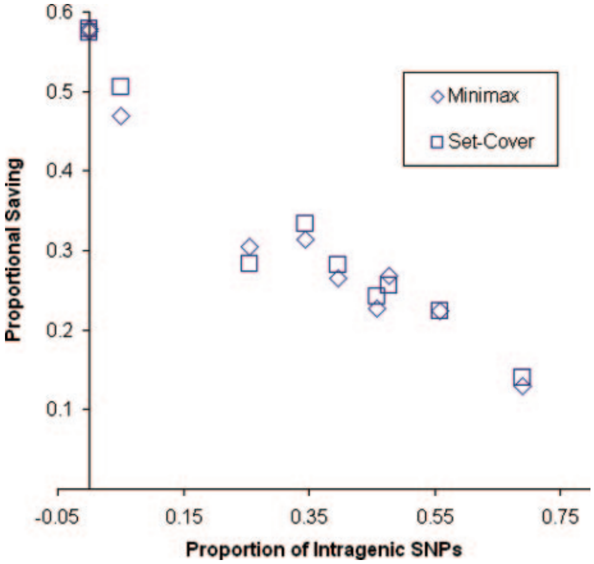

Fig. 1. Proportional saving in the number of selected tag SNPs that results from setting variable as against fixed tagging thresholds in the 10 ENCODE regions, plotted against the proportion of SNPs that are intragenic in these regions. Proportional saving is defined as (U − W)/U, where U is the number of tag SNPs selected based on a uniform tagging threshold of 0.8, while W is the number of tag SNPs selected based on a tagging threshold of 0.8 for intragenic SNPs and 0.4 for other SNPs.

In summary, WCLUSTAG allows users to prioritize different SNPs and genomic regions in a systematic association screen, depending on current genomic and disease data budget. The online web interface also permits users to import their own genotype data, or to directly withdraw HapMap data from the mirror database, for the calculations. A further area for development includes adding the facility for automatic query of genomic data in order to set tagging thresholds. The overall effectiveness of the tagging strategy will depend on the comprehensiveness of SNP maps, the quality of functional annotation of the genome, and the genetic architecture underlying complex human disease. Such factors remain to be explored in future studies.

ACKNOWLEDGEMENTS

The ENCODE genotype data were downloaded from HAPMAP site www.hapmap.org on Jun 30, 2004, based on NCBI build 34. The gene coordinates data were for Ensembl genes from NCBI build 34, downloaded from the UCSC Genome Browser http://genome-test.cse.ucsc.edu/cgi-bin/hgTables choosing human genome and July 2003 assembly. The work is supported by grants from the Research Grant Council of the Hong Hong SAR, China (HKU7579/05/M, HKU7669/06M, 7035/04P and 7035/05P), the National Eye Institute USA (EY-12562), the University of Hong Kong Strategic Research Theme on Genomics, Proteomics and Bioinformatics, Hong Kong Baptist University Faculty Research Grants, and Postgraduate Studentships (A.S.I. and K.J.S.H.) from the University of Hong Kong. Funding to pay the Open Access publication charges was provided by the Center for Mathematical Imaging and Vision, Hong Kong Baptist University.

Conflict of Interest: none declared.

REFERENCES

Ao,S.I. *et al*. (2005) CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics*, **21**, 1735–1736.

Byng,M.C. *et al*. (2003) SNP subset selection for genetic association studies. *Ann. Hum. Genet*., **67**, 543–556.

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

de Bakker,P.I. *et al*. (2005) Efficiency and power in genetic association studies. *Nat. Genet*., **37**, 1217–1223.

Halldorsson,B.V. *et al*. (2004) Optimal selection of SNP markers for disease association studies. *Hum. Hered*., **58**, 190–202.

Johnson,G.C. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet*., **29**, 233–237.

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.