<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Combining functional and linkage disequilibrium information in the selection of tag SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Sham, PC; Ao, SI; Kwan, JSH; Kao, P; Cheung, F; Fong, PY; Ng, MK</td>
</tr>
<tr>
<td><strong>Citation</strong></td>
<td>Bioinformatics, 2007, v. 23 n. 1, p. 129-131</td>
</tr>
<tr>
<td><strong>Issued Date</strong></td>
<td>2007</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://hdl.handle.net/10722/54299">http://hdl.handle.net/10722/54299</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>Bioinformatics. Copyright © Oxford University Press.; This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.; This is a pre-copy-editing, author-produced PDF of an article accepted for publication in Bioinformatics, following peer review. The definitive publisher-authenticated version Bioinformatics, 2007, v. 23 n. 1, p. 129-131 is available online at: <a href="http://bioinformatics.oxfordjournals.org/">http://bioinformatics.oxfordjournals.org/</a></td>
</tr>
</tbody>
</table>
Combining Functional and Linkage Disequilibrium Information in the Selection of Tag SNPs

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

¹Department of Psychiatry, ²Genome Research Centre, ³Department of Mathematics, University of Hong Kong, Hong Kong
⁴SGDP Centre, Institute of Psychiatry, King’s College London, UK
⁵Department of Mathematics, Hong Kong Baptist University, Hong Kong

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

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, on the basis 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 near-by 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 such as assay quality and whether the SNP has been genotyped. For example, if a coding SNP is given a weight of 0.8, and another non-coding SNP is given a weight of 0.4, then marker i can serve as a tag SNP for marker j.

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 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. For example, 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 R² 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 j to object i 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:

• A user-defined value of C is provided for each marker.
• The distance from marker i to marker j is defined as CᵢCᵢRᵢⱼ², where Cᵢ is the value of C specified for marker i. If CᵢCᵢRᵢⱼ²<0, then marker i can serve as a tag SNP for marker j.
• This asymmetric distance matrix is subjected to the minimax clustering method with the cut-off merging distance set at 0. 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 \( R_{ij} \). 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 \( C_{ij} R_{ij} \leq 0 \) 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 HapMap 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 33K 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% to 60% in the 10 ENCODE regions, depending on the proportion of the SNPs in the region that are intragenic (Figure 1).

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/hgTable choosing human genome and July 2003 assembly. The work is supported by grants from the Research Grant Council of the Hong Kong SAR, China (HKU7579/05/M and HKU7669/06M), 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 (ASI, KJSH) from the University of Hong Kong.

REFERENCES

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

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.