Vol. 19 no. 2 2003
Pages 287–288

Efficient selective screening of haplotype tag SNPs

Xiayi Ke* and Lon R. Cardon

Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK

Received on August 2, 2002; revised on September 6, 2002; accepted on September 11, 2002

ABSTRACT

Abstract: Haplotypes defined by common single nucleotide polymorphisms (SNPs) have important implications for mapping of disease genes and human traits. Often only a small subset of the SNPs is sufficient to capture the full haplotype information. Such subsets of markers are called haplotype tagging SNPs (htSNPs). Although htSNPs can be identified by eye, efficient computer algorithms and flexible interactive software tools are required for large datasets such as the human genome haplotype map. We describe a java-based program, SNPtagger, which screens for minimal sets of SNP markers to represent given haplotypes according to various user requirements. The program offers several options for inclusion/exclusion of specific markers and presents alternative panels for final selection.

Availability: The www-based program is available at http://www.well.ox.ac.uk/∼xiayi/haplotype/index.html.

Contact: xiayi@well.ox.ac.uk.

Haplotype-based methods offer a powerful approach to disease gene mapping, based on association between causal mutations and the ancestral haplotypes on which they arose (Gabriel *et al.*, 2002). Both regional (Jeffreys *et al.*, 2001) and chromosome-wide studies of linkage disequilibrium (LD) and haplotype structures (Patil *et al.*, 2001; Dawson *et al.*, 2002) have revealed blocks of limited haplotype diversity in which the majority of population samples can be characterized by only a few common haplotypes. Using the knowledge of these common haplotypes and the reduced sets of SNP markers that uniquely identify, or ‘tag’, them has the potential to greatly reduce the scale and cost of genotyping. These tagging SNPs have been called htSNPs by Johnson *et al.* (2001) who also provided a STATA script to aid their identification. While appropriate for the original dataset, many applications have specific needs that require further flexibility, such as the need to include or exclude specific SNPs (e.g. based on availability of PCR primers or on past performance of genotyping assays), the ability to review complete listings of all possible tag sets under different conditions, and the need to tag SNPs in exceptionally large datasets.

Here we describe an efficient tool that incorporates some of this flexibility and provides users with convenient access and use. The program is aided by a simple web interface in which users can set various options. For example, users can ignore uncommon haplotypes by setting a ‘coverage value’ less than 1.0, the default value, which identifies htSNPs for all observed haplotypes. Also, switches are available to force inclusion and exclusion of specific markers and inclusion of specific haplotypes. Moreover, the program provides support for haplotype data containing missing allele calls, which is an essential feature for large scale haplotype projects based on either traditional genotyping methods (Dawson *et al.*, 2002) or haploid cells (Patil *et al.*, 2001). The number of markers in any htSNP set ranges from 1 to T, but there is often more than one set of htSNPs having the same number of markers. They can be displayed by setting the ‘number of output sets’ in the interface. Certain flexibility is also allowed in the haplotype data input to accommodate haplotype and frequency estimation using different programs.

Our algorithm is based on the following set recovery process. Consider a matrix, P, containing $i = 1, \dots N$ haplotypes (rows) and $t = 1, \dots T$ markers (columns).

(1) For all pairs of haplotypes i and j $(i <> j)$, set $a_{ij}^{(t)} = 1$ if the allele at marker t differs between i and j; i.e. $P_{i,t} <> P_{j,t}$;

(2) Let $x^{(t)} = \begin{cases} 1 & \text{if marker t is included in the htSNP set} \\ 0 & \text{otherwise;} \end{cases}$

(3) Minimise $\sum_t x^{(t)}$subject to the constraint $\sum_t a_{ij}^{(t)} x^{(t)} \geqslant 1$ for all pairs i and j $(i <> j)$. A minimum number of htSNPs needed for any htSNP set is calculated as follows: for m haplotypes find the minimum n satisfying $2^n \geqslant m(n = \text{ceil}[\ln(m)/\ln(2)]$. A set recovery process is then employed to produce htSNPs sets

[*] To whom correspondence should be addressed

Downloaded from https://academic.oup.com/bioinformatics/article/19/2/287/372750 by U.S. Department of Justice user on 17 August 2022

with n or more htSNPs. For each htSNP set, the haplotype diversity captured is measured: if the set uniquely identifies two haplotypes, the htSNP set's diversity score is incremented by one. This measurement terminates at any time if the set fails to identify any pair of haplotypes, in which case, the set is not considered further.

Before applying the set recovery calculation, the following operations are conducted.

(1) Haplotypes are sorted according to their frequencies (descending). If a coverage value less than 1.0 is set by the user, the haplotypes are selected one by one into a separate haplotype set (the working haplotype set) until the cumulative haplotype frequency reaches the required coverage value. This process favors common haplotypes, i.e. it always takes the next most common of the remaining haplotypes.

(2) Users can bypass or overwrite the above haplotype selection process by explicitly indicating which haplotypes are included (via 'Rare haplotypes to be included'). All the specified haplotypes have the highest priority and are always selected into the working haplotype set even if their cumulative frequency exceeds the required coverage setting. If the required coverage is not met, Operation 1 is followed to select the remaining most common haplotypes. This option is particularly useful in combination with an appropriate coverage setting. It allows the explicit inclusion of a rare haplotype that is a slight variant of a common haplotype. Once a working haplotype set is established, all haplotypes in the set are treated equally regardless of their frequencies. Therefore, any candidate htSNP set should uniquely identify all haplotypes in the working set.

(3) Identify markers having identical patterns in the working haplotype set and keep only the first one of them since the others do not contribute additional distinguishing information (a pattern is considered identical if the Hamming distance between column vectors p_s and p_t, $H(p_s, p_t) = 0$, ignoring all haplotypes with missing data).

(4) All remaining markers are ranked according to their haplotype diversity values. These values are calculated by counting the number of major and minor allele appearances (e.g. in SNPs, '1's and '2's with '0's ignored) in the column, separately, and choosing whichever is smaller. These columns (markers) are then rearranged in such a way that the set recovery technique will start constructing a new htSNP set using markers with the highest diversity value possible.

(5) If there are markers specified to be included/excluded, they are checked against each other for any contradiction and/or redundancy. 'Included' markers have the highest priority and are always included into any htSNP sets. If more markers are needed, they are selected according to Operation 4, provided that they are not a member of the 'excluded' marker sets.

Johnson *et al.* selected htSNPs using haplotype diversities calculated for all possible htSNP sets whereas in our case the calculation is completed for only those that satisfy the haplotype coverage required by the user. Also, the present algorithm favors common haplotypes. In the exhaustive solution provided by Johnson *et al.* algorithm, several rarer haplotypes could contribute equally to a more common haplotype in terms of haplotype diversity. One of the consequences of our algorithm is, therefore, simplification of its htSNP set selection and calculation process because more controls are delegated to the user. This simplification reduces computation, thereby enabling the algorithm to handle potentially larger and more complicated haplotypes blocks.

ACKNOWLEDGEMENTS

The authors wish to thank the Wellcome Trust for support.

REFERENCES

Dawson,E., Abecasis,G.R., Bumpstead,S., Chen,Y., Hunt,S., Beare,D.M., Pabial,J., Dibling,T., Tinsley,E., Kirby,S. *et al.* (2002) A first-generation linkage disequilibrium map of human chromosome. *Nature*, **418**, 544–548.

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.

Jeffreys,A.J., Kauppi,L. and Neumann,R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.*, **29**, 217–222.

Johnson,G.C., 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 commondisease genes. *Nature Genet.*, **29**, 233–237.

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.

Downloaded from https://academic.oup.com/bioinformatics/article/19/2/287/372750 by U.S. Department of Justice user on 17 August 2022