Software for tag single nucleotide polymorphism selection

Daniel O. Stram

Division of Biostatistics and Genetic Epidemiology, Keck School of Medicine, University of Southern California, 1540 Alcazar Street, Room 220G, Los Angeles, CA 90033, USA
Correspondence to: Tel: +1 323 442 1817; Fax: +1 323 442 2349; E-mail: stram@usc.edu

Abstract

This paper reviews the theoretical basis for single nucleotide polymorphism (SNP) tagging and considers the use of current software made freely available for this task. A distinction between haplotype block-based and non-block-based approaches yields two classes of procedures. Analysis of two different sets of SNP genotype data from the HapMap is used to judge the practical aspects of using each of the programs considered, as well as to make some general observations about the performance of the programs in finding optimal sets of tagging SNPs. Pairwise R^2 methods, while the simplest of those considered, do tend to pick more tagging SNPs than are strictly needed to predict unmeasured (non-tagging) SNPs, since a combination of two or more tagging SNPs can form a prediction of SNPs that have no direct (pairwise) surrogate. Block-based methods that exploit the linkage disequilibrium structure within haplotype blocks exploit this sort of redundancy, but run a risk of over-fitting if used without some care. A compromise approach which eliminates the need first to analyse block structure, but which still exploits simple relationships between SNPs, appears promising.

Introduction

This review first describes the theoretical basis for single nucleotide polymorphism (SNP) tagging and then evaluates freely available software for the selection of tag SNPs for genetic association studies. The review is motivated by the author's experience as part of an active group of epidemiologists and statisticians and, especially, because of a now almost five-year-old ongoing collaboration between his group at the University of Southern California and members of the Broad Institute (formerly the Whitehead Institute) at MIT and Harvard University. The author has been involved in developing and implementing methods for selecting tag SNPs in candidate gene studies. In these studies, he and his colleagues have relied upon their own 'haplotype discovery' panel, in which, at present, almost 3,000 SNPs over approximately 30 genes in the steroid hormone pathway have been genotyped in 349 subjects in five ethnic groups (data publicly available at http://www.uscnorris.com/Core/DocManager/DocumentList.aspx?CID = 13).

These data have been used to: (1) identify the haplotype block structure of each of the candidate genes of interest; (2) estimate haplotype frequencies within blocks; and (3) use these haplotype frequency estimates to select informative non-redundant SNPs for use as tag SNPs in association studies.

For the purposes of this review, tag SNP methods are named according to the general approach described above as 'haplotype block-based'. As will be discussed below, block-based methods in which haplotype structure is first determined allow for the calculation of optimality of a set of tag SNPs in ways that are particularly relevant to certain types of analyses. Full determination of haplotype structure is not required by all proposed methods, however, and to a greater or lesser extent these may be regarded as 'haplotype block free'.

With the advent of the HapMap project (and the very recent release of a complete first pass of the project), over 1 million SNP genotypes, covering the entire human genome in 270 subjects sampled from four different ethnic groups, are now available for haplotype discovery, and thus may be considered for the use of tag SNP selection. At the time of writing, the HapMap project itself does not recommend any particular choice of SNPs for tagging, and this review is focused on the steps needed at this time in order to use HapMap data for the selection of tag SNPs using a number of different methods and software that is now available.

Key to the block-based method of selecting tag SNPs is to be able to estimate haplotype frequencies for large numbers of SNPs within blocks. Initial versions of the expectation-maximization (EM) algorithm used for estimating haplotype frequencies had severe restrictions in terms of the number of

SNPs that could be handled (since 2^n haplotypes are possible among n SNPs). The partition ligation EM algorithm (PLEM)[1] solves this problem, at least approximately, by performing the EM algorithm separately for small numbers of SNPs in each partition and then performing a restricted EM algorithm to merge (ligate) the partitions to form full haplotype frequency estimates. This approach allows the number of SNPs to be considered for haplotype frequency estimation to be greatly extended, provided that the total number of haplotypes in the region remains restricted (as in a haplotype block). Even the PLEM algorithm, however, becomes very slow, unwieldy and unreliable when there are large numbers of haplotypes to estimate. If there is no recombination between SNPs in a genomic region, there will be a maximum of n + 1 distinct haplotypes evident, but within large blocks, the number of haplotypes is typically much lower than this.[2] In regions with high rates of historical recombination, there comes an explosion in the number of haplotypes and, therefore, poor performance of the algorithm both in terms of reliability (especially where the total number of subjects genotyped is relatively small) and speed. Software programs that implement a PLEM or similar algorithm for calculation of haplotype frequencies include Haploview (see below for URLs), SNPHAP, Hapblock,[3] Haplore,[4] the latest version of TagIT[5] and the author's own program, tagSNPs.[6] Once haplotype frequencies are estimated (and subsequently treated as known), the basic idea for tag SNP selection is that a criterion function of the haplotype frequency distribution is decided upon and tag SNPs optimising that criterion are chosen. Optimising a given criterion function can involve a fairly lengthy calculation (the space of possible choices for tag SNPs being of size 2^n-1), and direct searches of all possible choices are not usually carried out. Instead, some kind of stepwise algorithm is employed. Examples of criteria functions are:

(1) Maximising the minimum R^2 between non-tag and tag SNPs (generally restricted to cover only the common unmeasured SNPs).

(2) Maximising the minimum R^2 between true haplotype count for a set of haplotypes h and their best prediction from the EM algorithm, with the h of interest generally restricted to only common haplotypes.

(3) Maximisation (or minimisation) of a global information-based estimate of haplotype diversity (involving both common and uncommon haplotypes).

In (1), the definition of R^2 can take three different forms, we may compute:

(1a) A bivariate pairwise R^2;

(1b) A multivariate allelic R^2;

(1c) A multivariate haplotype-based R^2.

The bivariate R^2 is the basis for a popular block-free approach (described below) to picking tag SNPs, so this review will focus on (1b) and (1c) for the time being. The difference between the multivariate allelic R^2 and multivariate haplotype R^2 is slightly subtle. The allelic R^2 is the true

value of the squared correlation that would be found in a linear regression of an unmeasured SNP s on tagging SNPs t_1, t_2,...t_{ms} in which only the main effects (counts) of the SNPs are allowed to enter the model (ie interactions of form $t_1 \times t_2$ etc are excluded as predictors). By contrast, the haplotype-based R^2 is the true value of the squared correlation between SNP s and the best prediction of SNPs from the haplotype estimates, which, under Hardy–Weinberg equilibrium (HWE) for the haplotypes, is of the form:

$$E\{\delta_s(H)|G\} = \frac{\sum_{H-G}\delta_s(H)p_{h_1}p_{h_2}}{\sum_{H-G}p_{h_1}p_{h_2}} \quad (1.1)$$

Where p_h is the frequency of haplotype h, $\delta_s(H)$ counts the number of copies (0, 1 or 2) of SNP s that is contained in the ordered haplotype pair H and the summation is over the set of ordered haplotype pairs for the SNPs that are consistent with the genotype data for the measured SNPs. Here, h and H refer to haplotypes that are made up of all SNPs (measured and unmeasured), whereas G refers only to the measured SNPs. The allelic R^2 and haplotype R^2 are formally equivalent (take the same value) when HWE for haplotypes holds and when there is no recombination between the SNPs considered. Otherwise, the allelic R^2 is lower (typically, only slightly lower) than the haplotype R^2 under HWE. If appropriate interaction terms (eg $t_1 \times t_2$) are included in the allelic R^2 calculation, this R^2 will increase, but its true value is still bounded above by the haplotype R^2, which is denoted R_s^2.

The haplotype R^2 in (2) above is different to that just described because it is now the haplotype h, rather than any single SNP, that is being predicted. Under HWE, the best prediction of haplotype h is of the same form as (1.1), with $\delta_s(H)$ replaced by $\delta_h(H)$, which counts the number of copies of haplotype h that are contained in the ordered haplotype pair H. The squared correlation, R_h^2 between $\delta_h(H)$ and its best prediction, $E\{\delta_h(H)|G\}$, is computed as described elsewhere.[6]

The information-based quantities referred to in (3) include such measures as $I_e = \sum_h p_h \log p_h$ (the information entropy) and $\sum_h p_h^2$ (the homozygosity index or probability that two randomly selected haplotypes are identical). For example, one might seek a set of tag SNPs of a given size that minimises the average posterior information entropy averaging over every possible configuration of measured genotypes G.

The haplotype and SNP R^2 criteria are specifically relevant to investigations in which either SNP-specific or haplotype-specific risks are being estimated in large cohort or case-control studies relating the occurrence of a disease of interest to the genomic variation in a particular candidate gene or gene locus. In particular, to a first approximation, R_s^2 and R_h^2 give the loss of information that results from not measuring a given SNP s or common haplotype h directly, compared with the analysis in which either SNP s was genotyped or haplotype.

h could be perfectly inferred. The global information-based quantities are less focused upon specific haplotypes and include both common and uncommon haplotypes in their definition of optimality.

Freely available computer programs for the selection of tag SNPs using one or more of the above-mentioned criteria include LDSelect, htSNP, Hapblock, TagIT, Tagger and tagSNPs.

Genuinely haplotype block-free approaches to picking tag SNPs are necessarily based upon simpler statistics than those described above. For example, the popular program, LDSelect, selects SNPs entirely upon pairwise R^2 between unmeasured SNPs and members of a tag SNP set. The benefit of this approach is that pairwise R^2 calculations are very reliable, irrespective of haplotype diversity, using haplotype discovery panels of a size similar to that of the HapMap or other typical studies. Thus, one may relatively safely provide large amounts of genotype data to the program with no prior consideration of haplotype block definition, and still expect reliable results.

Because the block-based approaches are, by definition, restricted to working within some sort of haplotype block definition (many have been considered in the literature), a problem arises; tag SNPs selected for one block may be partially or even wholly redundant with the tag SNPs selected in another block. Further, one is faced with the problem of how best to deal with regions of the gene of interest in which the particular definition of haplotype block one is working with is not satisfied. This and other issues will be discussed below.

Software specifics

The following freely available computer programs were evaluated: the block-based programs htSNP (in combination with SNPhap for the estimation of haplotype frequencies), TagIT, Hapblock and tagSNPs. In addition, the SNPs marked with triangles in the haplotype plots produced by the program Haploview (which the Haploview documentation calls tag SNPs, with no description of a specific method of choosing them provided) were evaluated as tag SNPs. Two block-free programs, LDSelect and Tagger, were also evaluated. All of these programs, except for Tagger, have versions available for Windows/DOS based computers which are downloadable from the internet (this author used his laptop computer running Windows XP in his evaluation). Tagger is an exception, in that it runs as a web application. Several of the programs ran not as 'native' executable programs, but rather as add-on programs to either Matlab or Stata programs. Since these are both popular and widely available (but not free), they were included in the author's evaluation; however, the author did not include certain other programs that required additional software (such as SAS Genetics) which he

had not already installed prior to writing this review. LDSelect is a Perl program, so a user may firstly have to install Perl; however, free versions of Perl are readily available for almost any computer.

Evaluation approach

The evaluation approach used here was first to investigate the completeness of features described in the documentation for each program considered and, secondly, to run the programs on two different SNP datasets (described below) using similar selection criteria to make a number of (as far as possible) comparisons between the tag SNP selections. The author worked mainly with the various R^2-based selection criteria described above; no single program could compute them all but most criteria could be used by at least two programs.

Some otherwise interesting programs were not included on the basis of missing features. For example, it is the author's belief that it is very important that the programs be able to include as tag SNPs those which are *a priori* identified as being of particular interest in an analysis. For example, it may be desirable to include a common missense SNP as a tag SNP or an SNP for which reports concerning its usefulness have already been published; forcing it in as a tag allows it to play a dual role, both as a candidate SNP itself and also contributing to defining the haplotype structure of the remaining SNPs. Programs that required that haplotype frequencies or other statistics (such as pairwise R^2) be computed externally were also avoided. (htSNP requires haplotype frequencies to be estimated separately using SNPhap; however, both are available from the same author, work together well and were considered as one program here.)

The author focused on the use of HapMap data (downloaded between 5th November, 2004 and 26th January, 2005) in his evaluation of the ease of use and capacity of each program, and used different genotype datasets in his evaluation. The first consisted of all the SNPs within the gene locus containing a specific candidate gene (*TGFBR1*), which in the author's estimation, had some 'typical' features of 'simple' genes (those with good haplotype block structure). Genotypes were downloaded (for the 30 CEPH Caucasian trios) for 15 common SNPs (frequency >3 per cent) in *TGFBR1* in a locus extending 20 kilobases (kb) upstream and 10 kb downstream of transcription of each gene. Using the default Gabriel rules as block definition, Haploview identified these SNPs as consisting of two blocks in *TGFBR1* (of three and ten SNPs in size, respectively), with two SNPs in no block. In addition to this candidate gene region, one of the ten densely genotyped ENCODE regions available from the HapMap website were used (again, for the 30 CEPH trios), specifically in order to determine how well the block-free programs would work when presented with a larger amount

of genotype data. The author used the ENM010 region (in chromosome band 7p15.2), in which, at the time of downloading the data, there were genotypes for 406 SNPs with a frequency ≥ 5 per cent.

Of the programs considered, only three were able to read the HapMap data directly. One of these was Haploview itself — which is now nicely incorporated into the HapMap website as a 'semi-official' way of looking at haplotype block structure. In addition, the web-based Tagger program and the author's tagSNPs program can both read the *.hmp formatted files obtained as HapMap genotype data dumps. Because no single format was read by all the programs being evaluated, the author spent some time adding features to the tag SNPs program, to write out the HapMap data in several various formats suitable for the other programs to read.

Table 1 provides a list of relevant features of the programs evaluated. Note that not all of the programs used were able to estimate haplotype frequencies directly from parent–offspring trios. For these programs, files were created containing only the genotypes for the 60 parents in the CEPH data. This implies some loss in the precision of estimation of haplotype frequencies, but rarely is this loss enough to have major effects on the SNPs chosen as tag SNPs.

Evaluation

Table 2 gives summary results for each of the programs applied to the data for the *TGFBR1* gene. In order to make a fair comparison of each program, the author attempted to use an as consistent as possible set of criteria for selecting tag SNPs over the set of seven programs evaluated:

(1) The block-free program LDSELECT was used to pick a set of tag SNPs so that the minimum value of pairwise R^2 between the measured and unmeasured SNPs would be equal to 0.9. Tagger was used for the same purpose, using both its pairwise and 'aggressive' mode (see below);

(2) For htSNP and TagIT, the criteria set for selecting SNPs included an R_s^2 of 0.9, this was used to select tag SNPs for each of the two blocks (SNPs 1–3 and SNPs 5–14), determined by Haploview separately;

(3) For tagSNPs, SNPs were chosen within each block using both an R_h^2 and an R_s^2 criteria of 0.90;

(4) For Hapblock, the 'block-finding' algorithm was set to be the empirical linkage disequilibrium (LD) method as close to the approach used by Haploview as possible, and used three methods (entropy, R_h^2 and pairwise R^2) to select tag SNPs;

(5) Using tagSNPs, the author calculated the R_s^2 and R_h^2 statistics that corresponded to the tagSNPs chosen by Haploview.

Some of the programs (LDSelect, Hapblock, TagIT) offered several equivalent choices for tag SNPs. When this was the

case, the author simply chose the first set listed for displaying in Table 2. Each of the programs correctly identified the redundancy between SNPs 1–3 in block 1 of the *TGFBR1* gene (SNPs 1 and 3 are perfectly correlated). For block 2, there was some difference between the programs. The most important difference is that the haplotype-based measures (R_h^2 and/or entropy) as computed by Hapblock or tagSNPs yielded one fewer tag SNP than did the R_s^2 criteria. Of course, it is reasonable that different criteria produce different numbers of tag SNPs. TagIT, htSNP and tagSNPs all yielded four SNPs (using R_s^2) to predict the ten total SNPs in block 2; in each case the R^2 values reached were equivalent (the minimum value found was 0.96).

It was also seen that, when using the pairwise R^2 criteria, both of the block-free programs, LDSelect and Tagger, selected the same number of tagging SNPs (eight) to cover the entire locus. When the Tagger program was set to its 'aggressive' option (basically a restricted R_s^2 calculation), it produced one fewer tag SNP. Hapblock also has a pairwise option, and also gave eight tagging SNPs based on pairwise R^2.

TGFBRI appears to have a considerable amount of recombination between block 1 and block 2 (multiallelic $D' = 0.13$). When all of the SNPs are considered, however, the total number of haplotypes is apparently relatively limited. For example, tagSNPs estimated that there were six common haplotypes (frequency > 5 per cent) that made up 90 per cent of the chromosomes over the entire locus. This apparent lack of diversity indicates that estimation of haplotype frequencies is probably sufficiently accurate, so that the block-based approaches can also be applied to the full gene in order to allow a more careful comparison between the block-based and block free programs. The results are given in Table 3.

Because of recombination between block 1 and block 2 (which causes an inherent lack of predictability in haplotype estimation), it was not possible to find tag SNPs that would meet the haplotype R_h^2 criterion > 0.90 for all of the common haplotypes over the whole gene. (Even using all of the SNPs gave a value of just 0.85 for one 7 per cent haplotype.) Therefore, the author restricted his comparison to SNPs picked using the R_s^2 criterion compared with the pairwise R^2. There is very little gain in efficiency using the multivariate (R_s^2) criterion compared with the pairwise, in that only one fewer tag SNP was identified (seven versus eight) by tagSNPs. Of the two other programs that compute R_s^2, one of them, TagIT, failed — apparently because its EM algorithm (which does not implement a partition ligation) could not easily handle all 15 SNPs (a newer version, TagIT 3.02, just released, evidently has remedied this deficiency and includes PLEM and TRIPLEM options). There was a slight discrepancy between tagSNPs and htSNP in the SNPs selected, although the same number was picked by each program. By the author's calculations, the SNPs selected by htSNP produced

Table 1. Features of the tag single nucleotide polymorphism (SNP) programs evaluated.

Program	Block-based?	Estimates blocks itself?	Reads HapMap data?	PLEM incorporates trio data?	Criteria for picking tag SNPs[a]	Allows 'force in' of special SNPs?	Platforms/required software
Haploview	YES	YES	YES	YES	Unknown	No	Java
Hapblock	YES	YES	NO	YES	Haplotype R_h^2, pairwise R^2, information entropy	Yes	Compiled C program for Win/DOS Unix and Linux
LDSelect	NO	N/A	NO	N/A	Pairwise R^2	YES	Perl
htSNP	YES	NO	NO	NO	Allelic R_s^2, haplotype R_s^b	NO	Stata + compiled C program (SNPHAP)
TagIT	YES	NO	NO	YES[b]	Allelic R_s^2, haplotype R_s^2	YES	Matlab
Tagger	NO	N/A	YES	YES	Pairwise R^2, restricted haplotype R_s^2	YES	Web application
tagSNPs	YES	NO	YES	YES	Haplotype R_h^2, haplotype R_s^2, pairwise R^2	YES	Compiled Fortran program for WIN/DOS Unix and Linux

[a] See text for definition of notation.
[b] Performs a standard EM rather than PLEM algorithm (see text for more information).

Table 2. Tag single nucleotide polymorphisms (SNPs) selected for *TGFBR1*.

Program	Block definition	Tag criteria	Tag SNPs in block 1, as defined by Haploview	Tag SNPs in block 2, as defined by Haploview	htSNPs not in blocks, as defined by Haploview
LDSelect	N/A	Pairwise R^2	1, 2	5, 6, 7, 13	4, 15
Haploview	Gabriel default[a]	Not stated	1, 2	5, 6, 7, 13	N/A
htSNP	Used results of Haploview	R_s^2	1, 2	5, 7, 8, 13	N/A
TagIT	Used results of Haploview	R_s^2	1, 2	5, 6, 13, 14	N/A
tagSNPs	Used results of Haploview	R_h^2	1, 2	5, 6, 13	N/A
		R_s^2	1, 2	5, 7, 8, 13	N/A
Hapblock	Empirical LD option[b]	R_h^2	1, 2	5, 9, 13	4
		I_e	1, 2	5, 7, 13	4
		Pairwise R^2	1, 2	5, 6, 7, 13	4, 15
Tagger	N/A	Pairwise R^2	2, 3	5, 6, 10, 13	4, 15
		Restricted R_s^2	2, 3	5, 10, 13	4, 15

[a] See reference 2.
[b] Hapblock found two blocks (SNPs 1–3, and SNPs 4–15).

a minimum R_s^2 of 0.9362, compared with 0.9952 for the SNPs selected by tagSNPs. Such a difference could either be due to the fact that the haplotype frequencies used by htSNP were

Table 3. Tag single nucleotide polymorphisms (SNPs) when all SNPs in *TGFBR1* are treated as being in one block.

Program	Tag criteria	htSNPs
LDSelect	Pairwise R^2	1, 2, 4, 5, 6, 7, 13, 15
Tagger	Pairwise R^2	2, 3, 5, 6, 10, 13, 15
	Restricted R_s^2	2, 3, 4, 5, 10, 13, 15
Haploview	Not stated	1, 2, 4, 5, 6, 13
htSNP	R_s^2	2, 3, 5, 10, 13, 14, 15
TagIT	R_s^2	Failed
tagSNPs	R_h^2	Criterion unreachable
	R_s^2	1, 2, 5, 6, 13, 14, 15
Hapblock	R_h^2	Criterion unreachable
	Entropy	1, 2, 4, 5, 6, 7, 13, 15
	Pairwise R^2	1, 2, 4, 5, 6, 7, 13, 15

different to those estimated using tagSNPs (since SNPhap does not use the trio information, whereas tagSNPs does), or because of the details of the search algorithm — the author used the stepwise down method in htSNP and the default forward selection (with a backwards substitution check) algorithm in tagSNPs. The htSNP program has an exhaustive search procedure but this is considered to be too slow for routine use. In any event, such a difference in R_s^2 is quite trivial.

The ENCODE (ENM010) region

The author was able to make the following comparisons using the 406 SNPs (with frequency ≥ 5 per cent) in the ENM010 region:
(1) Between the pairwise calculations in LDSelect and Tagger;
(2) Between the 'aggressive' SNP picking approach of Tagger and the pairwise approach;
(3) Between the above two and a 'relaxed' block-based approach, using Haploview to define the blocks and then extending the boundaries of the nearest neighbouring blocks to include those SNPs that were not included within blocks. TagSNPs, was used to pick SNPs based on R_s^2 separately for each block and pseudoblock.

For (1), the pairwise calculations using LDSelect identified that a total of 129 tag SNPs (31 per cent of all SNPs) were needed to achieve an R^2 of at least 0.9 between each unmeasured SNP and a single tag SNP. The same calculations performed by Tagger required 147 tag SNPs (36 per cent) in order to reach the same nominal criteria. In comparison (2), Tagger chose 93 tag SNPs (23 per cent) using its aggressive (partial R_s^2) mode. For comparison (3), Haploview found 27 blocks which contained the large majority of SNPs, so that at a maximum, an additional two SNPs were added to any one block when these boundaries were relaxed. The 'relaxed' block approach required 109 tag SNPs (27 per cent) to reach the minimum $R_s^2 \geq 0.9$ criteria. Haploview gave 123 tag SNPs as tagging SNPs for the same set of 'relaxed' blocks (but with unspecified tagging criteria).

General comments

At this point, it is not completely clear which criteria are most appropriate for the selection of tag SNPs. In the author's experience, there is little practical difference in picking tag SNPs within a block using either the R_h^2 or R_s^2 criteria,[7] in that both sets of criteria tend to suggest nearly equivalent SNPs at any given coverage level. (This makes sense because in order to predict SNPs using the haplotype approach, one must also be able to predict haplotypes.) A larger issue is whether simple pairwise methods are better than using the multivariate R_s^2 or R_h^2 methods. Pairwise methods inherently will choose more tag SNPs than will R_s^2 and this can be seen in the examination of the ENM010 region. The rationale for the use of pairwise R^2 methods appears to be an assumption that the goal of the statistical analysis in an association study using tag SNPs is to reproduce the results that would be achieved if every single known (measured or unmeasured) SNP was to be used as a predictor in the analysis. This is a laudable goal, but it can also be achieved by typing only the SNPs identified by the R_s^2 (or R_h^2) procedure and predicting them from formula (1.1) (this is implemented in tagSNPs). These predictions will be just as accurate as those derived from using the single SNPs identified in the pairwise procedure. Moreover, in any kind of multiallelic test (either haplotype-or-SNP-based), the multivariate measures of R^2 are a more directly relevant estimate of the actual amount of information contained in the tag SNP set.

The primary problem with consistently using the multivariate correlation measures for SNP selection is that restricted haplotype diversity is required in order to avoid over-fitting (giving an upwardly biased correlation estimate). By restricting the calculation of R_s^2 to be performed only for SNPs within blocks, the problem of over-fitting is largely eliminated. Block definitions are inherently quite arbitrary,

however, and there is often considerable redundancy between tag SNPs selected in two (or more) neighbouring blocks. The method that Tagger uses (in its aggressive mode) appears to be promising by simultaneously avoiding both over-fitting and redundancy. The Tagger output is useful because it gives information concerning which combinations of measured SNPs are required to reproduce which unmeasured SNPs. LDSelect gives this same information for the pairwise comparisons (by showing bins of similar SNPs). It is unclear why Tagger, in its pairwise algorithm, required more SNPs than did LDSelect (147 versus 129). This may be less important than its apparent improved performance in the aggressive mode over both pairwise and simple (relaxed) block-based methods.

Most of the programs were fairly easy to use and appeared to perform reasonably well when carrying out their designated tasks. Hapblock, however, proved problematic in two regards. First, it was by far the least user-friendly in term of requirements for user input — with a parameter file of exactly 11 lines, with all options indicated by difficult-to-remember numerical codes. Secondly, it proved impossible to get the program to work on the full ENM010 region — apparently because of limitations either in the number of SNPs or the size on an individual block. (The program would run for a while and then grind to a halt after processing about 80 SNPs, with no indication that it would ever complete.)

Tagger is a web-based program with a simple but reasonably general user interface, which allows both block and block-free operations. The author's own program, tagSNPs, has an extensive command language, allowing a user to perform numerous side calculations (forcing in and keeping out SNPs, picking a set of tag SNPs on the basis of one criterion — for example, R_s^2 — and checking its performance on the basis of another, R_h^2 or pairwise R^2, etc). As described above, this program is also able to predict both unmeasured SNPs and haplotypes involving unmeasured SNPs on the basis of the tag SNPs. The prediction part of this program is used for the implementation of such methods for case-control analysis as haplotype-specific risk estimation using standard logistic regression software. An interface for Statistical Analysis System (SAS, Cary, NC) has also been developed and is available on the author's website (http://www-ref.usc.edu/stram/tagSNPs.html).

For users already very familiar with either Matlab or Stata, the TagIT and htSNP programs may be attractive for picking tag SNPs. TagIT has an extensive set of commands and features and should be competitive now that it has implemented a PLEM algorithm) SNPhap (used by htSNP) would benefit from being able both to directly read the HapMap data dump files and utilise the trio data in the same way as is currently possible using tagSNPs, Tagger and Haploview.

Acknowledgments

URLs for cited software (last accessed 28th February, 2005)

Haploview http://www.broad.mit.edu/mpg/haploview/index.php

htSNP and SNPHAP http://www-gene.cimr.cam.ac.uk/clayton/software/

Hapblock http://www.cmb.usc.edu/msms/HapBlock/

Haplore http://zhao.med.yale.edu/softwarelist.html

tagSNPs, http://www-rcf.usc.edu/ ~ stram/tagSNPs.html

Tagger, http://www.broad.mit.edu/mpg/tagger/

TagIT, http://popgen.biol.ucl.ac.uk/software.html

LDSelect http://droog.mbt.washington.edu/ldSelect.html

References

1. Qin, Z.S., Niu, T. and Liu, J.S. (2002), 'Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms', *Am. J. Hum. Genet.* Vol. 71, pp. 1242–1247.
2. Gabriel, S.B., Schaffner, S.F., Nguyen, H. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science* Vol. 296, pp. 2225–2229.
3. Zhang, K., Qin, Z., Chen, T. *et al.* (2004), 'HapBlock: Haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms', *Bioinformatics* Vol. 12, pp. 2679–2692.
4. Zhang, K., Sun, F. and Zhao, H. (2005), 'HAPLORE: A program for haplotype reconstruction in general pedigrees without recombination', *Bioinformatics* Vol. 21, pp. 90–103.
5. Weale, M.E., Depondt, C., Macdonald, S.J. *et al.* (2003), 'Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping', *Am. J. Hum. Genet.* Vol. 73, pp. 551–565.
6. Stram, D.O., Haiman, C.A. and Hirschhorn, J.N. (2003), 'Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study', *Hum. Hered.* Vol. 55, pp. 27–36.
7. Stram, D.O. (2004), 'Tag SNP selection for association studies', *Genet. Epidemiol.* Vol. 27, pp. 365–374.