Genetics and population analysis

Genome-wide selection of tag SNPs using multiple-marker correlation

K. Hao

Algorithm and Data Analysis, Affymetrix, Inc., 3420 Central Expressway, Santa Clara, California, USA

Received on May 21, 2007; revised on September 8, 2007; accepted on September 28, 2007

Advance Access publication November 15, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivations: The tag SNP approach is a valuable tool in whole genome association studies, and a variety of algorithms have been proposed to identify the optimal tag SNP set. Currently, most tag SNP selection is based on two-marker (pairwise) linkage disequilibrium (LD). Recent literature has shown that multiple-marker LD also contains useful information that can further increase the genetic coverage of the tag SNP set. Thus, tag SNP selection methods that incorporate multiple-marker LD are expected to have advantages in terms of genetic coverage and statistical power.

Results: We propose a novel algorithm to select tag SNPs in an iterative procedure. In each iteration loop, the SNP that captures the most neighboring SNPs (through pair-wise and multiple-marker LD) is selected as a tag SNP. We optimize the algorithm and computer program to make our approach feasible on today’s typical workstations. Benchmarked using HapMap release 21, our algorithm outperforms standard pair-wise LD approach in several aspects. (i) It improves genetic coverage (e.g. by 7.2% for 200K tag SNPs in HapMap CEU) compared to its conventional pair-wise counter-part, when conditioning on a fixed tag SNP number. (ii) It saves genotyping costs substantially when conditioning on fixed genetic coverage (e.g. 34.1% saving in HapMap CEU at 90% coverage). (iii) Tag SNPs identified using multiple-marker LD have good portability across closely related ethnic groups and (iv) show higher statistical power in association tests than those selected using conventional methods.

Availability: A computer software suite, multiTag, has been developed based on this novel algorithm. The program is freely available by written request to the author at ke_hao@merck.com

Contact: ke_hao@163.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

It has been estimated that the human genome harbors >5 million common SNPs with minor allele frequency (MAF) of at least 10% (Carlson et al., 2003; Gonzalez-Neira et al., 2006; Kruglyak and Nickerson, 2001), and 7.5 million common SNPs with MAF of at least 5% (Barrett and Cardon, 2006). These polymorphisms explain a portion of the heritable risk for perhaps many diseases. There are two common strategies for constructing the contents of SNP genotyping panels, (1) SNPs chosen approximately randomly across the genome ignoring linkage disequilibrium (LD) patterns, and (2) LD-based tag SNPs chosen to maximize genetic coverage (Barrett and Cardon, 2006; Pe’er et al., 2006). Here, the genetic coverage is defined as the fraction of the set of all common (MAF ≥ 5%) SNPs exceeding some correlation threshold with at least one SNP typed by the array. The tag SNP approach takes advantage of our recent understanding in human genome’s fine LD structure and reduces genotyping costs (Carlson et al., 2003, 2004; Gonzalez-Neira et al., 2006). Driven by such a large potential benefit, a variety of algorithms have been proposed to efficiently identify tag SNPs, which is essentially a feature selection problem from the machine-learning viewpoint.

The SNP tagging strategy is tightly linked to the downstream testing methods for genetic association. If the selection starts from phased haplotype data and tag SNPs are picked to maximize the haplotypes they can distinguish, the downstream association studies might be more powerful when employing haplotype-based tests (Hao et al., 2005; Howie et al., 2006; Sebastiani et al., 2003). If the selection starts from diploid genotypes and tag SNP panels are developed to maximize genetic coverage through pair-wise LD (e.g. r²), single locus association testing could be more appropriate (de Bakker et al., 2005; Pe’er et al., 2006). Extending the pair-wise LD, r² among multiple markers is proposed to further increase the genetic coverage of SNP panels (de Bakker et al., 2005; Hao et al., 2006; Pe’er et al., 2006). For example, using combinations of two genotyped SNPs, the additional coverage gain is more than 10% for Illumina HumanHap300K panel in Caucasian (Pe’er et al., 2006). Software packages have also become available to quickly compute multiple-marker r² and achieve good coverage at a genome-wide scale (Barrett et al., 2005; Hao et al., 2006). It is noteworthy that such additional coverage gain is achieved on tag SNP panels that are developed solely using pair-wise r². How about selecting tag SNPs by incorporating multiple-marker LD information? In this article, we propose an extension of Carlson’s greedy algorithm (Carlson et al., 2004). Our new method identifies tag SNPs by simultaneously considering their pair-wise and multiple-marker LD with nearby neighbors. Furthermore, we evaluate the (1) gain in
2 METHODS
2.1 Data
The HapMap release 21 (2003) samples comprise 270 individuals from four populations: (1) 30 trios from the Yoruba, in Ibadan, Nigeria; (2) 30 trios from the CEPH collection (Utah residents with ancestry from Northern and Western Europe); (3) 45 unrelated Han Chinese individuals from Beijing, China and (4) 45 unrelated individuals from Tokyo, Japan. The Han Chinese and Japanese are often considered as a single East Asian sample (Gonzalez-Neira et al., 2006). The HapMap Project genotyped more than 4 million SNPs, among which about 2.2 million SNPs are common (MAF > 5%), and this number varies depending on the ethnic group.

2.2 Multiple-marker correlation
2.2.1 A SNP’s correlation with another marker ($r^2_{one}$) Consider SNP2 and its neighbor (SNP3) within a specified distance (e.g. 100kb), we term their pair-wise $r^2$ as $r^2_{one}$ because only a single SNP is used as predictor in tag SNP selection and downstream association test. Also, the notation of $r^2_{one}$ is consistency with multiple-marker $r^2$ notations.

2.2.2 A SNP’s correlation with another two markers ($r^2_{two}$) Herein, we implement a previously proposed method in computing multiple-SNP $r^2$ (de Bakker et al., 2005; Hao et al., 2006; Pe’er et al., 2006). Let us consider SNP3 and its two neighbors (SNP1 and SNP2) within certain distance. Each SNP carries two possible alleles (SNP, carries alleles A and a, SNP2 carries B and b, and SNP3 carries C and c). A multiple-marker $r^2$ can be used to quantify the correlation between SNP1 and the combination of SNP1 and SNP2. This combination of SNP1 and SNP2 may form four possible haplotypes (AB, Ab, aB and ab). Therefore, this SNP combination can be treated as a multi-allelic marker, which carries four alleles, denoted as AB, Ab, aB and ab. Pooling {AB, Ab and ab}, we transform this multi-allelic marker to a bi-allelic marker, which carries alleles AB and non-AB. We compute the pair-wise $r^2$ between this new bi-allelic marker and SNP3, and record the result as $r^2_{AB}$. Similarly, we calculate $r^2_{Bb}$ by pooling {AB, Ab and ab}. The same for $r^2_{Ab}$ and $r^2_{ab}$. Finally, we define $r^2_{two}$ between SNP3 and combination of SNP1 and SNP2; as max $2 \{ r^2_{AB}r^2_{AB}r^2_{ab}r^2_{ab} \}$.

2.2.3 A SNP’s correlation with another three markers ($r^2_{three}$) There are four SNPs (SNP1, SNP2, SNP3 and SNP4), and we are interested in $r^2_{three}$ of SNP4 with its three neighbors (SNP1, SNP2 and SNP3). SNP1, SNP2 and SNP3 form $2^3 = 8$ possible haplotypes. Again, we construct a novel bi-allelic marker by pooling seven haplotypes together, and obtain the $r^2_{three}$ after eight iterations. Similarly, $r^2_{three}$ or even high order of LD can be computed in the same framework.

2.3 Algorithm for selecting tag SNPs using multiple-marker LD (multiTag)
Notation:
(1) $S_{candidate}$: the set of candidate SNPs, from which tag SNPs will be selected. At the beginning of tag SNP selection, all SNPs belong to $S_{candidate}$.
(2) $S_{tagSNP}$: the set of tag SNPs. At the beginning of the selection, $S_{tagSNP}$ is empty, and it increases by one during each selection loop.
(3) $S_{captured}$: the set of SNPs already captured by $S_{tagSNP}$. At the beginning of the selection $S_{captured}$ is empty, and it increases during the tag SNP selection procedure.

Step 1, Initialization:
(1) We compute pair-wise $r^2 (r^2_{two})$ between every two SNPs in $S_{candidate}$ that are within a certain distance, $L$ (L defines the sliding window size. In practice, we usually set $L = 100,000$ or $200,000$ bp). In $S_{candidate}$, every SNP’s ability to capture its neighboring SNPs (by single-marker LD, $r^2_{one}$) is quantified with the SNP Capture Score (SCS).

$$SCS = \sum_{Sliding\ window} I_{(r^2_{one} > T_{one})}$$

Step 2, Iteration:
(1) We update the SCS for all remaining SNPs in $S_{candidate}$. For example, the $j$th SNP in $S_{candidate}$ denoted as candidate$_j$, its SCS can be calculated as (the following formula is limited to $r^2_{one}$ for illustration purposes, but the algorithm can readily accommodate $r^2_{three}$):

$$SCS = \sum_{candidate} I_{(r^2_{one} > T_{one})} + \sum_{additional\ member\ of\ candidate} I_{(r^2_{three} > T_{three})}$$

In the first $\Sigma$, we count how many members in $S_{candidate}$ (within candidate’s neighborhood including itself) are captured by candidate, through $r^2_{one}$. This computation is similar to that in Step 1, although with a smaller $S_{candidate}$. In the second $\Sigma$, we count additional members in $S_{candidate}$ that are not covered by candidate, through $r^2_{three}$. Here, $T_{three}$ is the threshold for ‘useful’ multiple-marker LD (e.g. $T_{three} = 0.9$).

(2) From $S_{candidate}$, we move the SNP with largest SCS to $S_{tagSNP}$ and denote it as tagSNP$_{1st}$ since it is the first member of $S_{tagSNP}$.

(3) From $S_{candidate}$, we move all captured SNPs as defined: a SNP as captured if it has an $r^2_{one} \geq T_{one}$ with tagSNP$_{1st}$ to $S_{captured}$.

Step 3, Termination:
We continue the iteration until (1) $S_{candidate}$ becomes empty, or (2) $S_{tagSNP}$ reaches a pre-specified size (e.g. 100,000 SNPs) or (3) the
Tag SNPs in CEU

Fig. 1. Genetic coverage of tag SNPs selected using (1) \( r_{\text{one}} \) only (dashed line), (2) incorporating \( r_{\text{two}} \) (solid line) and (3) further incorporating \( r_{\text{three}} \) (dotted line). The tag SNP selection and coverage calculation was conducted on HapMap release 21 data using a sliding window width of 100 kb.

Table 1. Number of tag SNPs \((\times 10^3)\) needed to achieve genetic coverage thresholds

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Algorithm</th>
<th>CEU</th>
<th>YRI</th>
<th>CHB-JPT</th>
<th>CHB</th>
<th>JPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>60%</td>
<td>One-marker*</td>
<td>88.4</td>
<td>266.5</td>
<td>86.5</td>
<td>83.3</td>
<td>80.2</td>
</tr>
<tr>
<td></td>
<td>Two-marker</td>
<td>66.7</td>
<td>197.6</td>
<td>69.9</td>
<td>66.4</td>
<td>63.5</td>
</tr>
<tr>
<td></td>
<td>Three-marker</td>
<td>65.9</td>
<td>189.1</td>
<td>68.9</td>
<td>65.0</td>
<td>62.2</td>
</tr>
<tr>
<td>80%</td>
<td>One-marker</td>
<td>203.6</td>
<td>637.8</td>
<td>192.8</td>
<td>185.4</td>
<td>178.6</td>
</tr>
<tr>
<td></td>
<td>Two-marker</td>
<td>143.5</td>
<td>426.6</td>
<td>146.4</td>
<td>138.0</td>
<td>131.3</td>
</tr>
<tr>
<td></td>
<td>Three-marker</td>
<td>137.6</td>
<td>400.2</td>
<td>142.4</td>
<td>133.5</td>
<td>126.2</td>
</tr>
<tr>
<td>90%</td>
<td>One-marker</td>
<td>356.4</td>
<td>888.4</td>
<td>329.6</td>
<td>318.5</td>
<td>305.5</td>
</tr>
<tr>
<td></td>
<td>Two-marker</td>
<td>234.7</td>
<td>670.8</td>
<td>243.7</td>
<td>221.1</td>
<td>209.9</td>
</tr>
<tr>
<td></td>
<td>Three-marker</td>
<td>225.7</td>
<td>657.0</td>
<td>234.7</td>
<td>212.8</td>
<td>201.2</td>
</tr>
</tbody>
</table>

*Equivalent to the standard Carlson’s greedy approach (Carlson et al., 2004).

 terse combinations of that are captured by independent validation sample set), we calculate the fraction of SNPs (dashed line), (2) incorporating \( r_{\text{two}} \) (solid line) and (3) further incorporating \( r_{\text{three}} \) (dotted line). The tag SNP selection and coverage calculation was conducted on HapMap release 21 data using a sliding window width of 100 kb.

coverage value reaches a prespecified level. Herein, the coverage can be easily calculated using the size of \( S_{\text{candidate}} \), \( S_{\text{tagSNP}} \) and \( S_{\text{captured}} \):

\[
\text{Coverage} = \frac{S_{\text{tagSNP}} + S_{\text{captured}}}{S_{\text{tagSNP}} + S_{\text{captured}} + S_{\text{candidate}}} \quad (3)
\]

For example, when \( S_{\text{candidate}} \) becomes empty the coverage is 100%.

2.4 Evaluation of genetic coverage and portability for tag SNPs

The genetic coverage on a training sample itself can be easily computed using formula 3, as presented in Figure 1. Recently, the portability of tag SNPs has attracted great interest, especially for populations from the same ethnic categories (e.g. Caucasian). For example, how well does \( S_{\text{tagSNP}} \) developed in HapMap CEU subjects collected from Utah perform on a Caucasian cohort collected in Europe? In this article, we look at portability among HapMap CHB and JPT cohorts. We identify tag SNPs in JPT (based on \( r_{\text{one}} \) and \( r_{\text{two}} \)), and record the two-SNP combinations of \( S_{\text{tagSNP}} \) members that contribute to SCS. In CHB (an independent validation sample set), we calculate the fraction of SNPs that are captured by \( S_{\text{tagSNP}} \) (aka, \( S_{\text{tagSNP}} \)'s genetic coverage in CHB) either by \( r_{\text{one}} \) or by recorded combinations of \( S_{\text{tagSNP}} \) members by \( r_{\text{two}} \).

3 RESULTS

To evaluate this novel multiple-marker SNP tagging approach, we apply it to HapMap release 21, which contains more than 4 million SNPs, a portion of which (e.g. ~2.1 million in CEU) are common SNPs, defined as MAF ≥ 5%. In this study, we only focus on these common SNPs (Barrett and Cardon, 2006; de Bakker et al., 2006). Figure 1 illustrates the tag SNP selection procedure with the thresholds \( T_{\text{one}} = 0.8 \), \( T_{\text{two}} = 0.9 \) and \( T_{\text{three}} = 0.95 \). For any \( S_{\text{tagSNP}} \) size, the tag SNPs selected using the multiple-marker approach (the solid line and dotted line) have higher coverage than \( r_{\text{one}} \) counterparts. The coverage improvement from \( r_{\text{one}} \) to \( r_{\text{two}} \) is sizable (e.g. 7.2% for 200 K tag SNPs in CEU). However, we observe limited additional gain when extending to \( r_{\text{three}} \). Suggesting combinations of three SNPs are less likely to be good surrogates of neighboring markers, at least for the MAF and \( T_{\text{three}} \) we considered. At the early phase of selection (when \( S_{\text{tagSNP}} \) is small), the slope of all three curves are substantial because tag SNPs are capturing large LD bins (Fig. 1). As the tag SNP increases, e.g. \( S_{\text{tagSNP}} \) in the interval \((1 \times 10^5, 2 \times 10^5)\), we observe gradually decreasing slope because smaller LD bins are being tagged. Furthermore, when \( S_{\text{tagSNP}} \) reaches ~276 K in CEU, the single-marker curve becomes linear, indicating that we have captured all LD bins and have started genotyping singleton SNPs. The refraction point comes earlier for multiple-marker tag SNPs. Conditioning on a fixed SNP number, multiple-marker tags offer higher coverage. From another viewpoint, the multiple-marker approach reduces genotyping costs for a given genetic coverage (Table 1). If we target 90% coverage in CEU, the single-marker algorithm requires 356.4 K SNPs, where the two-marker algorithm requires only 234.7 K tag SNPs, which translates into a 34.1% savings. Again, we observe only minor additional savings when extending to \( r_{\text{three}} \).

Tag SNPs optimized on a training dataset may not perform equally well on an independent study cohort, which was not used for tag SNP selection. Such a phenomenon is often referred to as portability loss, describing the genetic coverage decrease when applying tag SNPs to an independent sample set. In this article, we examined the portability of multiple-marker tag SNPs on two closely related ethnic populations (HapMap CHB and JPT). For example, in Figure 2A, we identified \( S_{\text{tagSNP}} \) using single-marker and two-marker approaches in JPT, and then evaluated \( S_{\text{tagSNP}} \)'s genetic coverage in JPT and CHB. It is noteworthy that, during two-marker tag SNP selection, we recorded the marker combinations that contribute to the tag SNP’s SCS. In the coverage calculation, only tag SNPs themselves and the recorded combinations were evaluated. By these means, the number of hypothesis tests only moderately
furthermore, two-marker $S_{\text{two}}$ has even higher genetic coverage in the validation sample than single-marker $S_{\text{one}}$ in the training sample.

In the context of HapMap release 21 (2.2 million common SNPs) and ±100 kb window size, about 108 calculations of $r^2$ are required for $r_{\text{one}}$ mode, because a given SNP has $n \approx 100$ neighbors within the window. For $r_{\text{two}}$ mode, because we examine all pairs of neighboring SNPs, the complexity goes from $C_n^1$ to $C_n^2$ (which translates to 100 times more calculations). The complexity goes even higher to $C_n^3$ for $r_{\text{three}}$ mode (Hao et al., 2006). Fortunately, number of calculations increases less dramatically in reality since many SNPs are captured at $r_{\text{one}}$ or $r_{\text{two}}$ mode and therefore no need to extend to higher order. Expertise in software development and programming is critical in its implementation. A GNU-licensed program suite, multiTag, is published along with this article and freely available. Written in ANSI C++, multiTag strongly emphasizes speed and scalability, and has been successfully tested on Windows XP, Linux and Sun Solaris platforms at a chromosome-wide scale. It is able to run in $r_{\text{one}}$ (equivalent to Carlson’s greedy method), $r_{\text{two}}$ or $r_{\text{three}}$ modes. Using HapMap release 21 CEU data and typical workstations (Intel Xeon 2.80 GHz CPU and 512 MB memory) as the test-bed, the $r_{\text{one}}$ mode finishes the entire genome within 1 h, in contrast, the $r_{\text{two}}$ and $r_{\text{three}}$ modes require ~100 and ~300 h, respectively, to finish a large chromosome (e.g. Chromosome 2). Fortunately, tag SNP selection on each chromosome can be run in parallel on a Linux cluster. If terminated prematurely (e.g. a Linux Cluster node crashes with unknown reason), multiTag is able pick up partial results and resume the computation, which appear to be a valuable feature when running the program for a long period. Incorporating $r_{\text{three}}$ is very computationally intensive, but may only yield small gain in genetic coverage (Fig. 1) or savings in genotyping (Table 1), suggesting tag SNP selection in $r_{\text{two}}$ order is more cost-effective.

4 DISCUSSION

Whole genome association study (WGAS) using tag SNPs is a powerful approach for elucidating genetic basis of common human diseases such as hypertension, type 2 diabetes mellitus and osteoporosis. A variety of techniques have been proposed in tag SNP selection (Barrett et al., 2005; Carlson et al., 2004; de Bakker et al., 2006; Halperin et al., 2005; Hao et al., 2005; Howie et al., 2006; Qin et al., 2006; Sebastiani et al., 2003; Stram et al., 2003), but many of them are only tested on relatively small chromosomal regions. Potentially, they can extend to genome-wide scale, although substantial modification is necessary to make them computational feasible in terms of memory usage and CPU run time. Because choosing tag SNP is literally a feature selection problem, several established feature selection algorithms were applied (Halperin et al., 2005; Horne and Camp, 2004; Lin and Altman, 2004; Phuong et al., 2005). However, these methods are still computationally complex, although not requiring exponential search time. As results, they can only be used on gene regions or small chromosomes. An alternative way is to focus on haplotype blocks, but the blocks are not always straightforward to define. Moreover, some feature selection methods (e.g. principal component analysis)

![Fig. 2. Summary of genetic coverage in portability experiments at the $r^2 \geq 0.8$ cutoff threshold. In the upper panel, we identify various number of tag SNPs in HapMap JPT data (i.e. the training dataset) using either single-marker approach or two-marker approach. These tag SNPs' genetic coverages are then evaluated in the training dataset itself (subject to over-fitting problem) and independent validation dataset (HapMap CHB). In the lower panel, we switch the training and validation datasets, and observe consistent results.](https://academic.oup.com/bioinformatics/article-abstract/23/23/3178/290076/3178.png?w=500)
derive mathematical abstractions, and mapping them to SNPs introduces one more level of complexity. Set theory has also been used (Sebastiani et al., 2003), but it only identify the perfect tag SNP sets (with 100% prediction power) and do not scale up to the entire genome. Currently, the block-free tag SNP selection strategy (Carlson et al., 2004) is employed by Illumina in developing whole-genome SNP arrays (Barrett and Cardon, 2006; Pe’er et al., 2006). These arrays are designed to optimize genetic coverage based on pair-wise $r^2 (r^2_{12})$. Interestingly, Pe’er and colleagues showed that by incorporating multiple-marker $r^2$ into the calculation, these SNPs offer more than 10% higher genetic coverage, which may also boost the statistical power of an association study (Hao et al., 2006; Pe’er et al., 2006). It is a natural question to ask, ‘could further gains in genetic coverage be achieved by incorporating multiple-marker $r^2$ as early as in the tag SNP selection step?” In this article, we propose a novel approach for tag SNP selection using multiple-marker $r^2$, and we systematically benchmark its performance. Our algorithm outperforms the standard Carlson’s approach in several aspects. (i) It improves genetic coverage (e.g. by 7.2% for 200 K tag SNPs in CEU) compared to the single-marker method, when conditioning on a fixed tag SNP number. (ii) It reduces genotyping costs when conditioning on a fixed genetic coverage (e.g. 34.1% savings in CEU at 90% coverage). (iii) Tag SNPs identified using our approach have a similar portability, as Carlson’s approach, across closely related ethnic groups (e.g. HapMap CHB and JPT collected in East Asia), and we believe this result can be generalized to cohorts of European ancestry.

Selecting a set of tag SNPs by exhaustive searching of all possible combinations is computationally intensive, and becomes impractical at the genome-wide scale even when limited to $r^2_{\text{one}}$ (Carlson et al., 2004; Hao et al., 2005; Qin et al., 2006; Sebastiani et al., 2003). When extending to the orders of $r^2_{\text{two}}$ and $r^2_{\text{three}}$, computation time and memory use become critical issues in algorithm development. Carlson’s greedy approach greatly reduces the search space (Hao et al., 2005), and therefore, is fast and memory efficient. The identified tag SNP set is fairly close to the minimum size although without a mathematical guarantee (Carlson et al., 2004; Howie et al., 2006; Qin et al., 2006). More importantly, a tag SNP set containing a certain degree of redundancy offers better portability than the mathematically minimal set (data not shown). Based on the above rational, we extend Carlson’s greedy method and elegantly incorporate higher order $r^2$ (e.g. $r^2_{\text{two}}$ and $r^2_{\text{three}}$). In each iteration, we only consider multiple-SNP $r^2$ formed by one candidate SNP and its neighbors in $S_{\text{tagSNP}}$, by these means, the search space is further reduced and the algorithm becomes computationally feasible. Shown in Figure 1, at the early phase of tag SNP selection, our approach (solid line and dotted line) is similar to Carlson’s method (dashed line), because $S_{\text{tagSNP}}$ is small and $r^2_{\text{two}}$ and $r^2_{\text{three}}$ make little contributions to SCS. As $S_{\text{tagSNP}}$ becomes larger, there are SNP combinations formed by a $S_{\text{candidate}}$ member and its neighbors in $S_{\text{tagSNP}}$ that give high $r^2_{\text{two}}$ or $r^2_{\text{three}}$. Hence, the genetic coverage of the multiple-marker algorithm starts to exceed Carlson’s method. It should be noted, when $S_{\text{tagSNP}}$ gets larger, more SNP combinations need to be evaluated in terms of $r^2_{\text{two}}$ and $r^2_{\text{three}}$, and the computational complexity grows quickly. Generally, there are two strategies in handling the large number of SNP combinations. (1) We could pre-compute all possible combinations’ $r^2_{\text{two}}$ and $r^2_{\text{three}}$ and using currently available software. These $r^2$ values are stored in either memory or hard disk, and then used in the SCS calculation during tag SNP selection. The drawback is the large memory requirement (if $r^2$ is stored in memory) or heavy file IO demand (if $r^2$ is stored in hard disk). (2) Alternatively, we can compute a given SNP-combination’s $r^2_{\text{two}}$ and $r^2_{\text{three}}$ on-the-fly. This strategy obviously has advantages in memory usage and/or file IO demand, however, more $r^2$ computation is required (because a certain SNP-combination’s $r^2$ value maybe used in several SCS calculations). The current version of multiTag employs the latter strategy, and therefore, can run on a typical workstation with 512 MB memory. The computation of $r^2_{\text{two}}$ and $r^2_{\text{three}}$ needs 3- and 4-SNP haplotype data, respectively. In our study, we directly used haplotypes (HapMap release 21) as input, which are reconstructed using the program PHASE (Marchini et al., 2006; Stephens et al., 2001). Our algorithm can accommodate diploid data, and reconstruct 3- or 4-SNP haplotypes on-the-fly, however, this strategy could be time consuming and potentially less accurate (Hao et al., 2006). As a result, researchers are recommended to first apply PHASE or other methods (Marchini et al., 2006) to accurately generate haplotypes, and then select tag SNPs using multiTag (the current version of multiTag only accommodates haplotype input).

In this study, we applied $T_{\text{one}} = 0.8$, $T_{\text{two}} = 0.9$ and $T_{\text{three}} = 0.95$ (Material and Methods section, Formulae 1 and 2). Certainly, we can choose different values for $T_{\text{two}}$ and $T_{\text{three}}$, e.g. a uniform $T_{\text{multiple}}$ (e.g. $T_{\text{two}} = T_{\text{three}} = 0.9$), which will not bias against three-marker tag SNPs. In the multiTag algorithm and computer software, these three threshold values ($T_{\text{one}}$, $T_{\text{two}}$ and $T_{\text{three}}$) can be flexibly tuned to achieve (1) differently sized $S_{\text{tagSNP}}$ (e.g. $S_{\text{tagSNP}}$ tends to be larger when higher $T$ values are applied); (2) various ratios between single-marker tag SNPs and multiple-marker SNPs and (3) various portability of resulting tag SNPs.

During SNP tagging, sometimes two or more candidate SNPs have equal SCS. In this situation, we randomly pick one of the best choices and continue the selection. Alternatively, we can formulate formula 1 to

$$SCS = \sum_{\text{Sliding Window}} I(r^2_{\text{two}} \geq r^2_{\text{one}}) \times f(r^2_{\text{one}})$$

where $f(r^2_{\text{one}})$ is a certain function of $r^2_{\text{one}}$, e.g. $f(r^2_{\text{one}}) = r^2_{\text{one}}$ or $f(r^2_{\text{one}}) = \sqrt{r^2_{\text{one}}}$. Because $r^2_{\text{one}}$ is a real number between 0 and 1, it is unlikely to observe a SCS tie when the training data has a reasonably large sample size. More importantly, such a modification biases towards SNPs in tighter LD with neighbors, hence it will further improve the tag SNPs’ average $r^2$ and portability.

Haploview has also implemented a multiple-marker tag SNP selection method (Barrett et al., 2005; de Bakker et al., 2006), but in a rather ad hoc manner. This algorithm works in two phases: (1) tag SNP selection based on pair-wise $r^2$, which is equivalent to Carlson’s greedy approach; (2) searching for specific multi-marker (haplotype) tests to improve tagging.
efficiency (de Bakker et al., 2006). The step (2) is done by iteratively dropping tag SNPs, one by one, and replacing them with a specific multi-marker predictor (using any of the remaining tag SNPs). That predictor is accepted only if it can capture the alleles originally captured by the discarded tag SNP; otherwise, that provisionally dropped tag is considered dispensable and kept (de Bakker et al., 2006). Obviously, this algorithm will miss some good two-marker predictors. For example, SNP1 is a single-marker tag for an LD bin and therefore recruited into $S_{\text{tagSNP}}$ by Haploview in phase (1). SNP2 by itself is a singleton, but the combination of SNP1 and SNP2 predicts a few other SNPs. Unfortunately, the Haploview algorithm will miss such a combination. To date, Haploview’s multi-marker tag SNP selection mode handles only about 10,000 SNPs (or ~10 Mb chromosome segment for HapMap release 21) in one run, and does not work at a chromosome-wide scale. Therefore, we did not conduct a head-to-head comparison between Haploview and multiTag.

Multiple testing remains as the primary challenge in WGAS. Many correction approaches have been proposed. (Bender and Lange, 2001; Chen et al., 2006; Hao et al., 2004; Herbert et al., 2006; Pe’er et al., 2006; Rosenberg et al., 2006; Wen et al., 2006) There are two strategies in dealing with multiple testing. (1) The statistical significance level should be adjusted by correction methods, and which method to apply depends on the nature of the SNPs being genotyped. For example, if the genotyped SNPs have weak LD among each other, Bonferroni correction would be adequate. (2) The number of hypotheses testing in WGAS should be carefully controlled. If we test all two or three marker combinations for genetic association with the study trait, the multiple comparison penalties may quick diminish statistical power. In this study, we record the marker combinations that contribute to genetic coverage (SCS) during tag SNP selection, and only these recorded combinations are tested for association in WGAS. By these means, we keep the number of testing in check. For example, in CEU, the multiple testing burden increases ~60% for 300 K two-marker tag SNPs comparing to 300 K single-marker tag SNPs (Fig. 3).

In term of statistical power, we investigate whether tag SNPs selected using multiTag (e.g. $r^2_{\text{two}}$ mode) outperform those selected using traditional approach (e.g. $r^2_{\text{one}}$ mode). In WGAS, it is difficult to calculate the absolute power directly. Fortunately, conditioning on false discovery rate (FDR), the number of discoveries reflects the relative power. Therefore, we can compare the power of tag SNPs by looking at number of discoveries at fixed FDR (5% and 10% in this article). The 90 HapMap Asian individuals are employed our power analysis, where we use 45 subjects (training sample) for tag SNP selection, and only these recorded combinations are compared to 300 K single-marker tag SNPs (Fig. 3).

Multiple testing remains as the primary challenge in WGAS. Many correction approaches have been proposed. (Bender and Lange, 2001; Chen et al., 2006; Hao et al., 2004; Herbert et al., 2006; Pe’er et al., 2006; Rosenberg et al., 2006; Wen et al., 2006) There are two strategies in dealing with multiple testing. (1) The statistical significance level should be adjusted by correction methods, and which method to apply depends on the nature of the SNPs being genotyped. For example, if the genotyped SNPs have weak LD among each other, Bonferroni correction would be adequate. (2) The number of hypotheses testing in WGAS should be carefully controlled. If we test all two or three marker combinations for genetic association with the study trait, the multiple comparison penalties may quick diminish statistical power. In this study, we record the marker combinations that contribute to genetic coverage (SCS) during tag SNP selection, and only these recorded combinations are tested for association in WGAS. By these means, we keep the number of testing in check. For example, in CEU, the multiple testing burden increases ~60% for 300 K two-marker tag SNPs comparing to 300 K single-marker tag SNPs (Fig. 3).

In term of statistical power, we investigate whether tag SNPs selected using multiTag (e.g. $r^2_{\text{two}}$ mode) outperform those selected using traditional approach (e.g. $r^2_{\text{one}}$ mode). In WGAS, it is difficult to calculate the absolute power directly. Fortunately, conditioning on false discovery rate (FDR), the number of discoveries reflects the relative power. Therefore, we can compare the power of tag SNPs by looking at number of discoveries at fixed FDR (5% and 10% in this article). The 90 HapMap Asian individuals are employed our power analysis, where we use 45 subjects (training sample) for tag SNP selection, and only these recorded combinations are compared to 300 K single-marker tag SNPs (Fig. 3).

Multiple testing remains as the primary challenge in WGAS. Many correction approaches have been proposed. (Bender and Lange, 2001; Chen et al., 2006; Hao et al., 2004; Herbert et al., 2006; Pe’er et al., 2006; Rosenberg et al., 2006; Wen et al., 2006) There are two strategies in dealing with multiple testing. (1) The statistical significance level should be adjusted by correction methods, and which method to apply depends on the nature of the SNPs being genotyped. For example, if the genotyped SNPs have weak LD among each other, Bonferroni correction would be adequate. (2) The number of hypotheses testing in WGAS should be carefully controlled. If we test all two or three marker combinations for genetic association with the study trait, the multiple comparison penalties may quick diminish statistical power. In this study, we record the marker combinations that contribute to genetic coverage (SCS) during tag SNP selection, and only these recorded combinations are tested for association in WGAS. By these means, we keep the number of testing in check. For example, in CEU, the multiple testing burden increases ~60% for 300 K two-marker tag SNPs comparing to 300 K single-marker tag SNPs (Fig. 3).

In term of statistical power, we investigate whether tag SNPs selected using multiTag (e.g. $r^2_{\text{two}}$ mode) outperform those selected using traditional approach (e.g. $r^2_{\text{one}}$ mode). In WGAS, it is difficult to calculate the absolute power directly. Fortunately, conditioning on false discovery rate (FDR), the number of discoveries reflects the relative power. Therefore, we can compare the power of tag SNPs by looking at number of discoveries at fixed FDR (5% and 10% in this article). The 90 HapMap Asian individuals are employed our power analysis, where we use 45 subjects (training sample) for tag SNP selection, and only these recorded combinations are compared to 300 K single-marker tag SNPs (Fig. 3).
(3) Kruskal–Wallis test is conducted between the trait and each tag SNPs (as well as the recorded tag SNP combinations). 
(4) We permute the trait value and repeat step (3) in order to derive FDR. Total 10,000 simulation loops are run, and we compare the relative power at FDR = 5 and 10% level (Fig. 4). Clearly, 200 K tag SNPs are more powerful than 100 K tag SNPs. More importantly, at fixed tag SNP number (or fixed genotyping cost), multiTag approach offers extra power even after adjusting for multiple testing. For example, at 10% FDR, 200 K tag SNPs derived in $r^2_{one}$ mode show 6.5% higher power than the counterpart of $r^2_{two}$ mode. 

Taken together, the tag SNP strategy is based on our recent understanding of the fine LD structure in the human genome. However, at the current stage, only the pair-wise LD information (e.g. $r^2$) is extracted. Herein, we present a novel approach, which utilizes the multiple-SNP LD in tag SNP selection. This approach efficiently reduces searching space, therefore, becomes computational feasible. Applied on HapMap release 21, multiTag uniformly outperforms traditional approaches in terms of both genetic coverage and statistical power, and we believe it will facilitate future genetic association studies.

ACKNOWLEDGEMENTS
The author wants to thank Dr. Joshua Millstein for insightful discussion and comments on the manuscript. The author also feels grateful to the reviewers for their valuable suggestions, which strengthen the paper greatly.

Conflict of Interest: none declared.

REFERENCES