European Journal of Human Genetics (2007) 15, 228–236
© 2007 Nature Publishing Group All rights reserved 1018-4813/07 $30.00
www.nature.com/ejhg

ARTICLE

Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance

Keyue Ding[1] and Iftikhar J Kullo*,[1]

[1]*Division of Cardiovascular Diseases, Mayo Clinic and Foundation, Rochester, MN, USA*

There is great interest in the use of tagging single nucleotide polymorphisms (tSNPs) to facilitate association studies of complex diseases. This is based on the premise that a minimum set of tSNPs may be sufficient to capture most of the variation in certain regions of the human genome. Several methods have been described to select tSNPs, based on either haplotype-block structure or independent of the underlying block structure. In this paper, we compare eight methods for choosing tSNPs in 10 representative resequenced candidate genes (a total of 194.2 kb) with different levels of linkage disequilibrium (LD) in a sample of European-Americans. We compared tagging efficiency (TE) and prediction accuracy of tSNPs identified by these methods, as a function of several factors, including LD level, minor allele frequency, and tagging criteria. We also assessed tagging consistency between each method. We found that tSNPs selected based on the methods Haplotype Diversity and Haplotype r^2 provided the highest TE, whereas the prediction accuracy was comparable among different methods. Tagging consistency between different methods of tSNPs selection was poor. This work demonstrates that when tSNPs-based association studies are undertaken, the choice of method for selecting tSNPs requires careful consideration.

European Journal of Human Genetics (2007) **15,** 228–236. doi:10.1038/sj.ejhg.5201755; published online 13 December 2006

Keywords: tagging SNPs; linkage disequilibrium; haplotype; association study

Introduction

Both linkage and association studies are employed to investigate the genetic basis of complex diseases such as atherosclerotic cardiovascular disease. Based on the premise that association mapping has greater power in identifying genetic determinants of complex diseases,[1] this study design is being increasingly used. Commonly, association studies test putative functional single nucleotide polymorphisms (SNPs) within candidate genes and regions – the so-called 'direct' approach.[2] An alternative is the 'indirect' approach (ie, a linkage disequilibrium (LD) – based approach), in which a subset of markers in a region of interest are selected from small panels of subjects and then used in large-scale association analyses.

The concept of LD is key to designing 'indirect' association studies for complex diseases. Regions with extensive LD, i.e., haplotype blocks, have been found interspersed with regions of medium and low LD in the human genome.[3,4] One way to reduce genotyping effort for association mapping of complex diseases is the use of haplotype tagging SNPs (htSNPs) or tagging SNPs (tSNPs). The two terms, htSNPs and tSNPs, refer to two different strategies for choosing the optimal minimum subset of SNPs from the entire set of SNPs. htSNPs are selected based on the haplotype-block model of LD pattern in a region of interest and represent the common haplotypes inferred from the original set of SNPs.[5] On the other hand, tSNPs are selected based on measures of association, such that a tSNP predicts partially or completely the state of other SNPs.[6]

*Correspondence: Dr IJ Kullo, Division of Cardiovascular Diseases, Mayo Clinic, 200 First Street Southwest, Rochester, MN 55905, USA.
Tel: +1 507 284 9049; Fax: +1 507 266 1702;
E-mail: kullo.iftikhar@mayo.edu

Received 22 December 2005; revised 7 September 2006; accepted 20 October 2006; published online 13 December 2006

Thus far, several methods to select htSNPs or tSNPs have been described, and these can be broadly classed into four categories. First, there are the methods based on defining how well a subset of SNPs captures the variation in the complete set.[5–11] The second category of methods is based on principal component analysis (PCA) to reduce the dimensions of complete sets of SNPs.[12,13] The third category is based on association or correlation between SNPs (ie, LD).[14,15] The fourth category includes several methods, an example being a method based on set theory that recursively searches the minimal set of SNPs with a given function.[16,17] These four categories can also be grouped into haplotype-block-based methods[5,7–9,12] and haplotype-block-free methods.[6,11,15] The tSNPs derived from these two classes give different genome coverage because of the varying 'blockiness' in the human genome.[18] In this paper, we use the term tSNPs to represent an 'optimal' selection of a subset of SNPs from the original set of SNPs, identified using either haplotype-block-based- or haplotype-block-free methods.

A recent review[19] described the methodological and conceptual differences between the available tagging algorithms. However, no systematic comparison of the available methods for selecting tSNPs has been performed, and a consensus method for choosing tSNPs has not been established. The researcher has been offered little guidance in the choice among these methods, that is, which method for choosing tSNPs is most appropriate for a particular candidate gene-based association study? We attempted to compare several leading tSNPs selection methods in 10 representative gene regions by using resequenced genotype data (pga.mbt.washington.edu). We assessed tagging efficiency (TE) and prediction accuracy of tSNPs derived by these methods. In addition, we investigated the impact of minor allele frequency (MAF) cutoff, tagging criteria, and LD level on tSNPs selection, as well as the tagging consistency between different methods.

Materials and methods

Gene selection

Sequence data for 87 candidate genes for atherosclerotic cardiovascular diseases from 23 European-Americans were downloaded from the SeattleSNPs database (http://pga.mbt.washington.edu) on March 11, 2005.[15,20] Ten representative genes (Table 1) were selected for comparing different tSNPs selection methods based on the following criteria: (1) LD level varied from strong LD ($D' > 0.8$), to moderate LD ($0.4 < D' \leq 0.8$), and to weak LD ($D' \leq 0.4$).[21] A measure of LD (D') was calculated using LDA software,[22] and the level of LD was assessed by use of sliding-window plots of average D' in each gene (Figure 1); (2) the length of the genes was close to the mean length of the 87 genes (mean ± SD of sequenced length in the 87 candidate genes = 21.3 ± 13.4 kb and median = 17.7 kb).

Table 1 LD level, number of SNPs, and singletons in 10 representative genes

Gene[a]	*LD level*	*Length (kb)*	*SNPs (n)*	*Singletons (n)*
DO	High	17.8	17	2
IL1A	High	18.4	47	3
MGP	High	9.9	11	0
VKORC1	High	11.2	13	3
ALOX12	Moderate	18.6	55	9
SELL	Moderate	24.8	95	23
VCAM1	Moderate	22.9	38	10
F2	Weak	22.1	52	9
F10	Weak	29.5	61	16
ICAM1	Weak	19.0	35	11

[a] Gene abbreviations: *DO*, dombrock blood group; *IL1A*, interleukin 1 α; *MGP*, matrix G1a protein; *VKORC1*, vitamin K epoxide reductase complex, subunit 1; *ALOX12*, arachidonate 12-lipoxygenase; *SELL*, selectin L; *VCAM1*, vascular cell adhesion molecule 1; *F2*, coagulation factor II; *F10*, coagulation factor X; *ICAM1*, intercellular adhesion molecule 1.

Methods for the selection of tSNPs

We compared eight published methods of identifying tSNPs. Most of the methods are based on searches to evaluate subsets of SNPs using different measures and include All common haplotypes,[7,8] Haplotype diversity,[5] Coefficient of determination (R^2_h),[10] Haplotype entropy (Entropy),[23] and Haplotype r^2 (TagIT).[11] Another set of methods is based on PCA, for example, the method described by Lin and Altman.[13] Carlson *et al*[15] developed the method LD r^2 (based on pairwise LD), in which the maximally informative site and all associated sites are grouped into a bin. Sebastiani *et al*[17] described a method (BEST) in which all optimum fully informative tSNPs are generated based on set theory. A brief summary of each method, including measures or statistics, comments and original references, is presented in Supplementary Table 1.

Haplotype-block definitions

The above eight methods can also be classified as haplotype-block-based methods (eg, All common haplotypes, Haplotype diversity, R^2_h, and Entropy) and haplotype-block-free methods (eg, TagIT, LD r^2, PCA, and BEST). Haplotype blocks have been defined based on diversity,[7,8] LD,[3] and recombination.[24] Comparisons of haplotype blocks based on these definitions have revealed similarity between the LD-based method and the recombination-based method.[25–27] We choose to define haplotype blocks based on LD when haplotype-block-based tSNPs selection methods were employed. The LD-based haplotype-block definition requires that the proportion of SNP pairs with strong D' (absolute $D' \geq 0.70$) must account for at least 95% of pairs of SNPs.[3]

Selection of tSNPs under different MAF and tagging criteria

The following programs were available from the authors' websites: Hapblock,[28] ldSelect for the method LD r^2,[15]

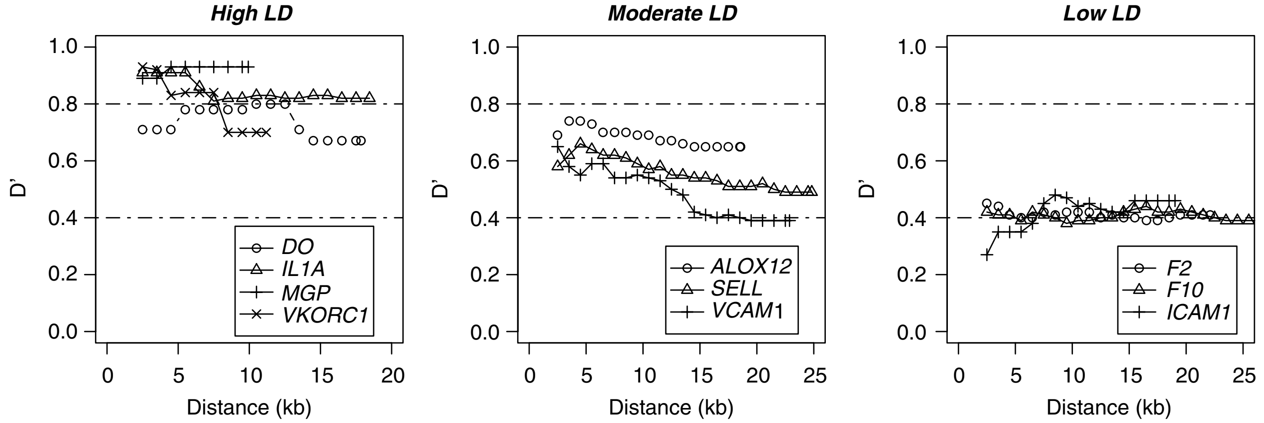

Figure 1 Sliding window plots of average LD measure (D') in the 10 genes included in the present study. Average D' was calculated from all SNP pairs in 5 kb sliding windows (1 kb increment between windows starting from 2.5 kb). If there were no SNPs in a given window, the D'-value was assigned as a value of 'not available'. Different LD levels were present in different gene regions: strong LD ($D' > 0.8$), moderate LD ($0.4 < D' \leq 0.8$) and weak LD ($D' \leq 0.4$).

TagIT,[11] and BEST.[17] Hapblock integrates several methods, including All common haplotypes, Haplotype diversity, R_{h}^2, and Entropy. We implemented the PCA algorithm[13] in MatLab®, using varimax-rotation method to map tSNPs after eigenSNPs were mathematically selected.

We used tagging criteria as a measure for quantifying the proportion of variation captured by a tSNPs set. The tagging criteria of 0.70, 0.80, and 0.90 at MAFs of 0.10 and 0.20 were assessed in six methods, including Haplotype diversity,[5] Entropy,[23] R_{h}^2,[10] LD r^2,[15] Haplotype r^2 (TagIT),[11,29] and PCA.[13] For each method, combinations of parameters of tagging criteria and MAF were input. For the remaining two methods, that is, All common haplotypes[7,8] and BEST,[17] there was no need to input the tagging criteria. We compared TE, prediction accuracy, and tagging consistency between different tSNPs selection methods as described below.

Tagging efficiency (TE)

TE was defined based on Ke *et al*[30] as

$$TE = n/n_h \tag{1}$$

where n_{h} is the number of tSNPs and n is the total number of SNPs under different MAF cutoffs. The measure of TE provides an estimate of the savings in genotyping offered by tSNPs and is expected to vary under different MAF cutoffs.

We also selected a 100 kb ENCODE region (ENCyclopedia Of DNA Elements, a project that aims to produce a dense set of genotypes across large genomic regions) on chromosome 7q21.13 (www.hapmap.org) to test the effect of sample size (ie, $n = 24$, 48, 72, and 90) on TE.

Prediction accuracy

Halperin *et al*[31] have proposed a measure of prediction accuracy to evaluate the quality of tSNPs and described its utility in selecting tSNPs given the genotype information of SNPs from a set of unrelated individuals. The measure aims to maximize the expected accuracy of predicting untyped SNPs, given the unphased information of the tSNPs.[31] Formally, for a given set of SNPs t, the objective is to find a set of tSNPs S of size t and a prediction function f such that the prediction error is,

$$\eta = \sum_{j=1}^{m} \Pr[f_j(Z_S(g)) \neq g(j)] \tag{2}$$

where, Z_S is the restriction of the genotype to the tSNPs position, and $g(j)$ is the j-th SNP in genotype g. We calculated the prediction accuracy of the sets of tSNPs generated by the eight methods under two MAF in 10 genes using the program Gerbilview.[32]

Tagging consistency

Let set$_1$ and set$_2$ denote two sets of tSNPs derived either from one population using two different methods or from two populations using one method. To quantify the consistency or similarity between the sets of tSNPs, we used the methods of Schwartz *et al*[27] and Liu *et al*[33] that assess whether or not tSNPs from two different methods or two populations coincide. The P-value ($P(set_1, set_2)$) is from Fisher's exact test for the null hypothesis that the two tSNPs sets are independent.

$$P(set_1, set_2) = \sum_{i=m}^{\min(B_1,B_2)} \frac{\binom{B_1}{i}\binom{L-B_1}{B_2-i}}{\binom{L}{B_2}} \tag{3}$$

where B$_1$, B$_2$ are the numbers of tSNPs in set$_1$ and set$_2$, respectively, m is the number of tSNPs shared by set$_1$ and set$_2$, and L is the total number of SNPs in the regions under study. The consistency measure (C) is defined as the negative logarithm of the $P(set_1, set_2)$ value,

$$C = -\log P(set_1, set_2). \tag{4}$$

Results

The eight methods of tSNPs selection were compared in 10 genes (Table 1). These 10 genes had different levels of LD and their genomic length was close to the average length of the 87 candidate genes for atherosclerotic cardiovascular disease. The LD pattern of these genes for European-Americans is illustrated in Figure 1, indicating strong LD in *DO*, *IL1A*, *MGP*, and *VKORC1*, moderate LD in *ALOX12*, *SELL*, and *VCAM1*, and weak LD in *F2*, *F10*, and *ICAM1*.

Tagging efficiency

For each method of selecting tSNPs, we defined TE as the total number of markers in the region of interest divided by the number of tSNPs chosen by a particular method, based on Eq. (1). The TE of the eight tSNPs selection methods across different gene regions with different levels of LD, using two MAF cutoffs (0.10 and 0.20) and tagging criterion of 0.90, is shown in Table 2 and Supplementary Figure 1. The mean TE varied (from ~ 2 to ~ 25) depending on the method of tSNPs selection and the LD level of the gene region. The overall TE was highest for Haplotype diversity and TagIT. As expected, the LD level in the gene regions affected TE; for most tSNPs selection methods, TE was higher in strong LD regions than in regions of moderate LD and weak LD. The variance in TE in the high LD regions was weakly related to method of tSNPs selection (two-way analysis of variance (ANOVA), $P = 0.090$) but not to MAF ($P = 0.998$). In the moderate and low LD regions, the variance in TE could be attributed to the method used for tSNPs selection ($P = 0.028$ and $P = 0.003$ in moderate and low LD regions, respectively) as well as MAF ($P = 0.002$ and $P = 0.009$ in moderate and low LD regions, respectively).

To investigate whether TE is significantly affected by tagging criteria under the two MAF, we performed ANOVA for six methods (no tagging criteria were input for methods of All common haplotypes and BEST; see Supplementary Table 1) under different MAF cutoffs and three different tagging criteria (0.70, 0.80, and 0.90). TE was significantly affected by tagging criteria, for several methods, especially PCA, TagIT, Entropy, and Haplotype diversity, in regions of high LD (Supplementary Table 2). In regions of moderate and low LD, TE was significantly affected by tagging criteria for the methods PCA and TagIT.

We also tested the effect of sample size (ie, $n = 24$, 48, 72, and 90) on TE in the 100 kb ENCODE region on chromosome 7q21.13. No significant difference in TE was noted at different sample sizes (Supplementary Figure 2).

Prediction accuracy of tSNPs

The prediction accuracy of tSNPs for gene regions with different LD levels, using the two MAF cutoffs (0.10. and 0.20) and tagging criterion of 0.90 is shown in Table 3 and Supplementary Figure 3. The prediction accuracy of tSNPs was comparable among different methods in gene regions with different LD levels. If we set prediction accuracy of 0.90 as a threshold in regions with high or moderate LD, the prediction accuracy under different MAF cutoffs exceeded or approached the threshold for all eight methods. There was no significant difference in prediction accuracy of tSNPs under the two MAF cutoffs in any of the eight methods. No significant differences in prediction accuracy were noted among different methods at various LD levels ($P > 0.05$, two-way ANOVA). Thus, neither the choice of tSNPs selection method nor level of LD affected the prediction accuracy.

Tagging consistency

In general, different methods for selecting tSNPs generated different sets of tSNPs. To quantify and examine the consistency between tSNPs generated by different methods of tSNPs selection, we used the tSNPs similarity measure in Eq. (4) and the test of significance in Eq. (3). We compared the results generated under a MAF cutoff of 0.10 and tagging criterion of 0.90. There was greater similarity between tSNPs identified by the methods All common haplotypes, Entropy, R_h^2, and BEST, than between the remaining methods. Supplementary Table 3 summarizes the statistical tests of the null hypothesis of independent tSNPs from pairwise comparison of the methods All common haplotypes, Entropy, R_h^2 and BEST. It can be seen that most of the pairwise comparisons among these four methods provided evidence against the null hypothesis of independent SNP selection by different methods ($P < 0.05$). When comparing regions with different levels of LD, we observed that the fraction of significant results of pairwise comparisons decreased with decreasing LD level. In case of *IL1A* (high LD), six out of six pairwise comparisons were significant compared with three out of six in the *ALOX12* gene (moderate LD), and two out of six in the *F2* gene (weak LD).

Computational cost

Finally, we compared the computational cost of each method of tSNPs selection. Five methods (All common haplotypes, haplotype diversity, Entropy, LD r^2, and R_h^2) were run under Linux with an AMD® athlon 2800 + MP CPU and the other three methods (BEST, PCA, and TagIT) were run under a Windows XP® system with a 2.8 GHz CPU. The computational cost was comparable at MAF 0.10 and 0.20 among the different methods. For example, in the *ALOX12* gene, with MAF = 0.10 and tagging criterion 0.90, R_h^2 and Entropy took < 10 min to get the results using the Hapblock program, whereas the other methods took < 1 min. However, when comparing tSNPs selection methods in larger genomic regions (such as the 500 kb ENCODE regions), the computational burden for the methods R_h^2, Entropy, and BEST was significant. For example, the runtime for the program BEST grew exponentially and we

Table 2 Tagging efficiency of the eight methods of tSNPs selection

		Methods							
Gene	*MAF*[a]	*Hap.Div*	*ACH*	*Entropy*	R_h^2	*TagIT*	*PCA*	*BEST*	*LD* r^2
DO	0.10	15.00	7.50	3.00	3.75	5.00	3.75	3.00	3.00
	0.20	12.00	11.00	6.00	12.00	12.00	4.00	6.00	6.00
IL1A	0.10	21.00	10.50	10.50	10.50	21.00	14.00	10.50	10.50
	0.20	20.50	13.67	13.67	13.67	20.50	13.67	13.67	10.25
MGP	0.10	11.00	5.50	3.67	5.50	5.50	5.50	2.75	3.67
	0.20	11.00	5.50	3.67	5.50	5.50	5.50	2.75	3.67
VKORC1	0.10	5.00	3.33	3.33	2.50	5.00	2.50	3.33	2.00
	0.20	5.00	3.33	3.33	2.50	5.00	2.50	3.33	2.00
ALOX12	0.10	10.25	5.86	4.56	5.13	20.50	10.25	4.10	2.93
	0.20	7.75	5.17	4.43	5.17	15.50	7.75	4.43	3.44
SELL	0.10	13.00	10.40	8.67	3.06	17.33	10.40	8.67	3.06
	0.20	4.25	2.83	2.43	2.13	5.67	3.40	3.40	1.42
VCAM1	0.10	3.00	1.91	1.91	2.10	4.20	3.00	2.10	1.91
	0.20	2.17	1.86	1.63	2.17	4.33	3.25	1.63	2.17
F2	0.10	6.00	3.60	4.50	3.60	6.00	6.00	3.60	3.60
	0.20	1.67	1.67	1.67	1.67	2.50	1.67	1.67	1.67
F10	0.10	2.38	1.72	1.72	1.35	6.20	3.44	2.82	1.48
	0.20	2.11	1.36	1.36	1.19	4.75	2.71	2.11	1.36
ICAM1	0.10	8.00	4.00	2.67	1.60	4.00	2.29	1.78	1.60
	0.20	4.33	4.33	2.60	1.63	4.33	2.60	2.17	1.86

Hap. Div: Haplotype diversity; ACH: All common haplotypes.

[a] Minor allele frequency.

Table 3 Prediction accuracy of the eight methods of tSNPs selection

		Methods							
Gene	*MAF*[a]	*Hap.Div*	*ACH*	*Entropy*	R_h^2	*TagIT*	*PCA*	*BEST*	*LD* r^2
DO	0.10	0.896	0.881	0.865	0.945	0.862	0.881	0.965	0.943
	0.20	NA[b]	NA	0.961	NA	NA	0.966	0.922	0.943
IL1A	0.10	0.984	1.000	1.000	1.000	0.984	0.783	1.000	0.991
	0.20	0.842	1.000	1.000	1.000	0.991	0.757	1.000	1.000
MGP	0.10	NA	0.957	0.946	0.976	0.976	0.976	0.981	0.951
	0.20	NA	0.947	0.946	0.976	0.976	0.976	0.981	0.951
VKORC1	0.10	0.853	0.634	0.634	0.746	0.832	0.725	0.634	0.696
	0.20	0.853	0.634	0.634	0.768	0.832	0.725	0.634	0.800
ALOX12	0.10	0.953	0.977	0.991	0.991	0.926	0.941	1.000	0.993
	0.20	0.924	0.981	0.995	0.989	0.946	0.970	1.000	0.987
SELL	0.10	0.791	0.864	0.716	0.858	0.764	0.800	0.755	0.919
	0.20	0.806	0.877	0.896	0.908	0.820	0.837	0.826	0.904
VCAM1	0.10	0.851	0.909	0.857	0.901	0.899	0.904	0.905	0.904
	0.20	0.938	0.986	0.965	0.925	0.939	0.957	0.965	0.932
F2	0.10	0.864	1.000	0.969	1.000	0.938	0.938	1.000	1.000
	0.20	1.000	1.000	1.000	1.000	0.929	1.000	1.000	1.000
F10	0.10	0.829	0.819	0.773	0.859	0.699	0.723	0.743	0.730
	0.20	0.817	0.800	0.800	0.986	0.713	0.641	0.700	0.757
ICAM1	0.10	0.848	0.717	0.622	0.855	0.877	0.908	0.621	0.964
	0.20	0.935	0.822	0.918	0.887	0.926	0.918	0.615	0.964

Hap. Div: Haplotype diversity; ACH: All common haplotypes.

[a] Minor allele frequency.

[b] Prediction accuracy cannot be calculated due to one tSNP.

were unable to get results even in 2 weeks for a 227 (marker number) by 180 (haplotype sample size) matrix under a windows server (2 GHz CPU and 3.50 GB of RAM).

Discussion

In this paper, we compared tagging efficiency (TE), prediction accuracy, and tagging consistency of tSNPs generated from eight published methods of tSNPs selection. The comparisons were carried out using sequence data for 10 representative candidate genes for atherosclerotic cardiovascular disease with varying levels of LD in a sample of European-Americans (Figure 1, Table 1).

Several factors, including LD level, MAF, tagging criteria, and sample size may affect TE. TE was affected significantly by the level of LD and was higher in genes with higher level of LD. In high LD regions, the amount of variance in TE was weakly related to the different methods of tSNPs selection but not MAF. However, in moderate and low LD regions, TE was influenced by the method of tSNPs selection as well as the MAF cutoff. There appeared to be nearly an order of magnitude difference in TE between some of the methods (eg, a lower efficiency using LD r^2 and a higher efficiency using TagIT for *ALOX12*) (Table 2 and Supplementary Figure 1). This difference may be due to long-range associations between SNPs. For example, LD may exist between bins, which were partitioned based on LD r^2 (or between haplotype blocks, such as All common haplotypes), whereas TagIT is able to incorporate such long-range LD.[3,29] Tagging criteria influenced TE in the genes with strong or moderate LD levels, especially for the methods PCA, TagIT, and Haplotype diversity (Supplementary Table 2). With increase in tagging criteria from 0.70 to 0.80 to 0.90, a greater number of tSNPs was needed to tag the entire gene region for these three methods. As the SeattleSNPs investigators used a relatively small number ($n = 23$ European-Americans) of subjects for SNP ascertainment, we evaluated whether the perceived TE was affected by larger sample sizes. No significant change in TE was noted using larger sample sizes (n ranged from 24 to 90) for each tSNPs selection method (Supplementary Figure 2).

The prediction accuracy of tSNPs selected by different methods approached or exceeded the threshold of 0.90 (Table 3 and Supplementary Figure 3). Neither the choice of tSNPs selection methods nor the level of LD significantly affected the prediction accuracy (two-way ANOVA, $P > 0.05$). Given the higher TE of Haplotype diversity and TagIT, the prediction accuracy of these two methods was higher in the gene regions with high LD and comparable to other methods in the moderate and low LD regions.

In order to investigate whether TE and prediction accuracy were different in genes larger than the ones we initially studied (10 genes, 10 ~ 30 kb), we assessed TE and prediction accuracy in an additional five genes ranging in size from 30 to 50 kb. A similar pattern for TE and prediction accuracy among different methods was noted (Supplementary Figure 4).

We plotted prediction accuracy (on the X-axis) *versus* TE (on the Y-axis) for the 10 genes for each tSNPs selection method to assess the tradeoff between prediction accuracy and TE at various LD levels (Supplementary Figure 5). However, for a given method of tSNPs selection, no simple linear relationship between TE and prediction accuracy was obvious in the 10 genes. We also calculated the measure 'tagging effectiveness' described by Ke *et al*[34] to assess the percent of hidden (untyped) SNPs with $r^2 \geq 0.80$ to the haplotypes defined by a tSNPs set (Figure 2). For all the eight tSNPs selection methods, tagging effectiveness in high LD regions was much higher than that in moderate and low LD regions. However, within the three levels of

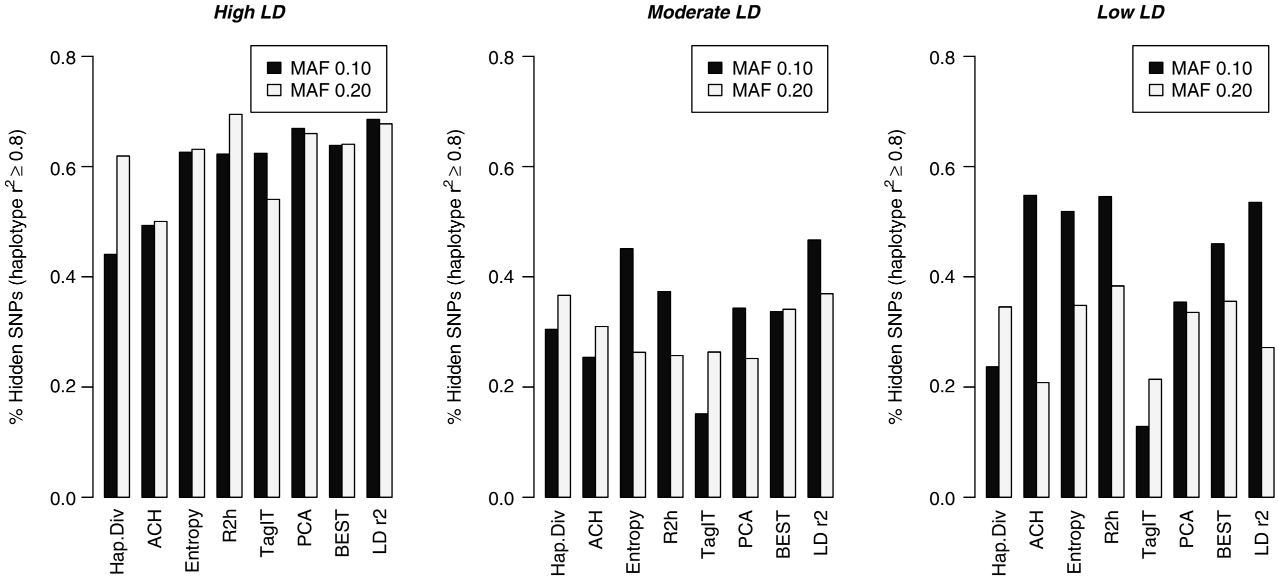

Figure 2 The percent of hidden (untyped) SNPs with $r^2 \geq 0.80$ to haplotypes defined by a tSNPs set under two MAF cutoffs.

LD, tagging effectiveness was similar among the eight tSNPs selection methods ($P > 0.05$ by ANOVA).

Pairwise comparisons of tSNPs sets revealed poor consistency between tSNPs selected using any two of the eight methods. A limited degree of tagging consistency was present between tSNPs derived from four methods (All common haplotypes, Entropy, R_h^2, and BEST), three of these methods (All common haplotypes, Entropy, and, R_h^2) being haplotype-block-based (Supplementary Table 3). This may be due to the low haplotype diversity in each block for the haplotype-block-based methods, and therefore a greater likelihood for similar tSNPs to be selected to represent common haplotypes using two different methods. Among haplotype-block-free methods, the underlying principles for choosing tSNPs are diverse; for example, TagIT incorporates all associations between SNPs along a region, whereas LD r^2 considers association between SNPs in a bin (Supplementary Table 1). Thus, the tSNPs sets identified by haplotype-block-free methods differed considerably. We found little similarity between tSNPs sets generated from the remaining four methods (Haplotype diversity, LD r^2, TagIT, PCA) (analyses not shown). Forton *et al*[35] have suggested that haplotype reconstruction by tSNPs generated by haplotype-block-based methods is more accurate than haplotype-block-free methods.

The International HapMap project is meant to facilitate the optimal selection of SNPs for cost-effective and robust whole-genome association studies.[36] Using the methods described above, we obtained tSNPs sets for the same 10 genes in 24 African-Americans using resequenced data from the SeattleSNPs database. Tagging consistency between European-Americans and African-Americans was measured using Eq. (3). We found that the tagging consistency between the two ethnic groups, or 'tSNPs transferability',[37] for any of the eight methods was poor, indicating that the tSNPs set selected for European-Americans are not transferable to African-Americans (analyses not shown). However, tSNPs may be transferable between different geographical samples of an ethnic group[37] or between various non-African populations.[38]

At least two initiatives, SeattleSNPs[15,20] and the Environmental Genome Project (EGP),[39] have resequenced several hundred candidate genes involved in inflammation and environmental response, to facilitate candidate-gene-based association studies. These two projects used a small panels of subjects ($n = 23$–30) belonging to different ethnic groups to characterize polymorphic variation and pattern of LD in the candidate gene regions. In the present study, we used resequenced data from SeattleSNPs ($n = 23$ European-Americans and 24 African-Americans). It has been estimated that 48 chromosomes would identify $\sim 99\%$ of SNPs with a MAF$\geq 5\%$.[40] In a simulation study, Thompson *et al*[41] found that using such a sample size (25 unphased individuals) to select tSNPs did not reduce the power of an association study, compared to using all SNPs.

Comparing various tSNPs selection methods is far from straightforward. First, selecting representative gene data sets for analysis is problematic because of different LD levels in different genes and the variability in the number of SNPs among genes. Second, the size of candidate genes and genomic regions to be studied could be much larger than the regions (50 kb maximum) investigated in the present study and the extent of LD could also extend well beyond this size. Third, there is no consensus on what are the most appropriate statistics to evaluate the performance of tSNPs sets. Each method for choosing tSNPs has its own quality measure to optimally select a set of tSNPs. The measure we used for evaluating the quality of tSNP selection – prediction accuracy – aims to maximize the expected accuracy of predicting untyped SNPs, given the unphased (genotype) information of the tSNPs.[31] Fourth, there is no simple relationship between TE and accuracy that allows one to choose an optimal balance of these two measures. Recently, Ke *et al*[34] used a matching TE among three tSNPs selection methods to assess 'tagging effectiveness'. Generating a matching TE to compare prediction accuracy of eight tSNPs selection methods would require significant computational resources beyond the scope of the present study.

A major expectation from using tSNPs is that the genotyping cost is reduced, whereas at the same time the statistical power for identifying associations is only minimally compromised. Statistical power may be an important metric in deciding which method is the most optimal in association studies. A direct comparison of tSNPs selection methods in the context of statistical power may be possible in a simulation study,[42] but was outside the scope of the present study. Another expectation of tSNPs selection methods is flexibility, allowing one to force a specific SNP, for example, a nonsynonymous SNP, into the set of tSNPs. Some programs, such as Hapblock,[28] allow insertion of a specific SNP into a tSNPs set. Flexibility would also allow one to replace a SNP that cannot be genotyped with an alternate tSNP, for example, an alternate SNP within the same bin (LD r^2) or the same haplotype block (Haplotype diversity).

Except for LD r^2, which uses genotype data to calculate the pairwise LD measure (r^2), the methods for selecting tSNPs are based on haplotype data. We used haplotypes inferred from the PHASE program[43] to generate the input for each method. Although convenient, statistical inference of haplotypes is associated with a degree of uncertainty as a proportion of the inferred haplotypes may be incorrect. This may reduce the statistical power of a haplotype approach to detect an association with disease.[43,44] How the tSNPs selection methods compare when genotype data is used instead of haplotypes needs further study. The use of genotype data combined with a PL–EM (Partitioning-Ligation–Expectation-Maximization) algorithm for choosing tSNPs maybe comparable to the use of haplotypes in association studies.[9]

A limitation of the present analyses is that there are moderate amounts of missing data in the 10 genes (missing data ranged from 1.9 to 8.2%). The PHASE program imputes missing data when haplotypes are constructed. How the missing data rate might affect tSNPs selection is unclear although Zhang *et al*[9] found that the statistical power and the number of tSNPs with and without moderate missing data were similar, even with 10% data missing.

In conclusion, our comparison of the performance of several methods for choosing tSNPs revealed that TE varied with the methods, being highest for Haplotype Diversity[5] and TagIT (haplotype r^2).[11] Because the prediction accuracy and the computational cost were similar among different methods, the methods Haplotype Diversity and TagIT may be considered initially for tSNPs selection. We found limited tagging consistency between tSNPs generated by different tSNPs selection methods, and tSNPs had limited transferability between African-Americans and European-Americans. This work demonstrates that when tSNPs-based association studies are undertaken, the choice of method for selecting tSNPs requires careful consideration.

Acknowledgements

This work was supported in part by NIH grant RR-17720 (I.J.K). We acknowledge the Supercomputing Institute of University of Minnesota, Minneapolis for technical support.

References

1 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
2 Peltonen L, McKusick VA: Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* 2001; **291**: 1224–1229.
3 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
4 Jeffreys AJ, Neumann R: Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* 2002; **31**: 267–271.
5 Johnson GC, Esposito L, Barratt BJ *et al*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.
6 Halldorsson BV, Bafna V, Lippert R *et al*: Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res* 2004; **14**: 1633–1640.
7 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.
8 Zhang K, Deng M, Chen T, Waterman MS, Sun F: A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* 2002; **99**: 7335–7339.
9 Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F: Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res* 2004; **14**: 908–916.
10 Stram DO, Haiman CA, Hirschhorn JN *et al*: Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* 2003; **55**: 27–36.
11 Weale ME, Depondt C, Macdonald SJ *et al*: Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am J Hum Genet* 2003; **73**: 551–565.
12 Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* 2003; **73**: 115–130.
13 Lin Z, Altman RB: Finding haplotype tagging SNPs by use of principal components analysis. *Am J Hum Genet* 2004; **75**: 850–861.
14 Wu X, Luke A, Rieder M *et al*: An association study of angiotensinogen polymorphisms with serum level and hypertension in an African-American population. *J Hypertens* 2003; **21**: 1847–1852.
15 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
16 Bafna V, Halldorsson BV, Schwartz R, Clark AG, Istrail S: Haplotypes and informative SNP selection algorithms: don't block out information. *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology* 2003. New York, USA: ACM Press, pp 19–27.
17 Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF: Minimal haplotype tagging. *Proc Natl Acad Sci USA* 2003; **100**: 9900–9905.
18 Wall JD, Pritchard JK: Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet* 2003; **73**: 502–515.
19 Halldorsson BV, Istrail S, De La Vega FM: Optimal selection of SNP markers for disease association studies. *Hum Hered* 2004; **58**: 190–202.
20 Crawford DC, Carlson CS, Rieder MJ *et al*: Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* 2004; **74**: 610–622.
21 Abecasis GR, Noguchi E, Heinzmann A *et al*: Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet* 2001; **68**: 191–197.
22 Ding K, Zhou K, He F, Shen Y: LDA – a java-based linkage disequilibrium analyzer. *Bioinformatics* 2003; **19**: 2147–2148.
23 Nothnagel M, Furst R, Rohde K: Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered* 2002; **54**: 186–198.
24 Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 2002; **71**: 1227–1234.
25 Ding K, Zhou K, Zhang J, Knight J, Zhang X, Shen Y: The effect of haplotype block definitions on inference of haplotype block structure and htSNPs selection. *Mol Biol Evol* 2005; **22**: 148–159.
26 Schulze TG, Zhang K, Chen Y-S, Akula N, Sun F, McMahon FJ: Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum Mol Genet* 2004; **13**: 335–342.
27 Schwartz R, Halldorsson BV, Bafna V, Clark AG, Istrail S: Robustness of inference of haplotype block structure. *J Comput Biol* 2003; **10**: 13–19.
28 Zhang K, Qin Z, Chen T, Liu JS, Waterman MS, Sun F: HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics* 2005; **21**: 131–134.
29 Goldstein DB, Ahmadi KR, Weale ME, Wood NW: Genome scans and candidate gene approaches in the study of common

diseases and variable drug responses. *Trends Genet* 2003; **19**: 615–622.

30 Ke X, Durrant C, Morris A *et al*: Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum Mol Genet* 2004.

31 Halperin E, Kimmel G, Shamir R: Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics* 2005; **21** (Suppl 1): i195–i203.

32 Kimmel G, Shamir R: GERBIL: genotype resolution and block identification using likelihood. *Proc Natl Acad Sci USA* 2005; **102**: 158–162.

33 Liu N, Sawyer SL, Mukherjee N *et al*: Haplotype block structures show significant variation among populations. *Genet Epidemiol* 2004; **27**: 385–400.

34 Ke X, Miretti MM, Broxholme J *et al*: A comparison of tagging methods and their tagging space. *Hum Mol Genet* 2005; **14**: 2757–2767.

35 Forton J, Kwiatkowski D, Rockett K, Luoni G, Kimber M, Hull J: Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms. *Am J Hum Genet* 2005; **76**: 438–448.

36 The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.

37 Mueller JC, Lohmussaar E, Magi R *et al*: Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* 2005; **76**: 387–398.

38 Huang W, He Y, Wang H *et al*: Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc Natl Acad Sci USA* 2006; **103**: 1418–1421.

39 Livingston RJ, von Niederhausern A, Jegga AG *et al*: Pattern of sequence variation across 213 environmental response genes. *Genome Res* 2004; **14**: 1821–1831.

40 Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001; **27**: 234–236.

41 Thompson D, Stram D, Goldgar D, Witte JS: Haplotype tagging single nucleotide polymorphisms and association studies. *Hum Hered* 2003; **56**: 48–55.

42 Zhang K, Calabrese P, Nordborg M, Sun F: Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet* 2002; **71**: 1386–1394.

43 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.

44 Zhang S, Pakstis AJ, Kidd KK, Zhao H: Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet* 2001; **69**: 906–914.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)