Genetic Epidemiology 28: 207–219 (2005)

Analysis of Single-Locus Tests to Detect Gene/Disease Associations

Kathryn Roeder[1]*, Silviu-Alin Bacanu[2], Vibhor Sonpar[2], Xiaohua Zhang[3] and B. Devlin[2]

[1]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*
[2]*Department of Psychiatry University of Pittsburgh, Pittsburgh, Pennsylvania*
[3]*Department of Biometrics Research Merck Research Laboratories West Point, Pennsylvania*

A goal of association analysis is to determine whether variation in a particular candidate region or gene is associated with liability to complex disease. To evaluate such candidates, ubiquitous Single Nucleotide Polymorphisms (SNPs) are useful. It is critical, however, to select a set of SNPs that are in substantial linkage disequilibrium (LD) with all other polymorphisms in the region. Whether there is an ideal statistical framework to test such a set of 'tag SNPs' for association is unknown. Compared to tests for association based on frequencies of haplotypes, recent evidence suggests tests for association based on linear combinations of the tag SNPs (Hotelling T^2 test) are more powerful. Following this logical progression, we wondered if single-locus tests would prove generally more powerful than the regression-based tests? We answer this question by investigating four inferential procedures: the maximum of a series of test statistics corrected for multiple testing by the Bonferroni procedure, T_B, or by permutation of case-control status, T_P; a procedure that tests the maximum of a smoothed curve fitted to the series of of test statistics, T_S; and the Hotelling T^2 procedure, which we call T_R. These procedures are evaluated by simulating data like that from human populations, including realistic levels of LD and realistic effects of alleles conferring liability to disease. We find that power depends on the correlation structure of SNPs within a gene, the density of tag SNPs, and the placement of the liability allele. The clearest pattern emerges between power and the number of SNPs selected. When a large fraction of the SNPs within a gene are tested, and multiple SNPs are highly correlated with the liability allele, T_S has better power. Using a SNP selection scheme that optimizes power but also requires a substantial number of SNPs to be genotyped (roughly 10–20 SNPs per gene), power of T_P is generally superior to that for the other procedures, including T_R. Finally, when a SNP selection procedure that targets a minimal number of SNPs per gene is applied, the average performances of T_P and T_R are indistinguishable. *Genet. Epidemiol.* 28:207–219, 2005. © 2005 Wiley-Liss, Inc.

Key words: haplotype; single nucleotide polymorphism; nonparametric regression; multivariate regression

The supplemental materials described in this article can be found at http://www.interscience.wiley.com/jpages/0741–0395/suppmat
Contract grant sponsor: NIH; Contract grant numbers: MH057881 and DA011922.
*Correspondence to: Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.
E-mail: roeder@stat.cmu.edu
Received 25 June 2004; Revised 9 August 2004; Accepted 18 September 2004.
Published online 6 January 2005 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20050

INTRODUCTION

It is now possible to genotype a large number of single nucleotide polymorphisms (SNPs) at relatively low cost per SNP and sample. The maturing of genotype technology coincides with the push to understand the genetic basis of complex disease. This fortuitous coincidence of events permits intense evaluation of candidate genes/linkage regions by genotyping a set of SNPs covering, in terms of linkage disequilibrium (LD), the entire region of interest.

The number of SNPs required to cover a genomic region depends on the LD structure of the region. Recent studies of LD in the human genome [Daly et al., 2001; Reich et al., 2001; Patil et al., 2001; Gabriel et al., 2002; Clark et al., 1998; McVean et al., 2004; Crawford et al., 2004] demonstrate that the structure of pairwise LD in the genome is quite variable, implying the number of SNPs required to cover regions is also highly variable. Indeed, neighboring polymorphisms can exhibit complete LD or none at all.

The nature of pairwise LD has inspired the investigation of higher level LD structure, such as that embodied by haplotypes [Reich et al., 2001; Abecasis et al., 2001; Daly et al., 2001; Maniatis et al., 2002]. When data from specific populations are analyzed, a substantial fraction of the genome demonstrates limited diversity of haplotypes, or

haplotype blocks. Recognition of this block structure in part underlies the conjecture that tests for association typically would be more powerful if they were haplotype-based, as opposed to a series of single locus tests, because haplotypes capture more of the LD structure and thus could be more efficient. Due to the extreme heterogeneity of LD structure, however, it seems unlikely that there is a single best approach to association testing. Indeed, some investigations illustrate that haplotype-based tests of association are more powerful than tests based on multilocus genotypes or a series of single-locus tests, while others reach the opposite conclusion [Zhang et al., 2002a; Chapman et al., 2003].

Two alternative strategies lie midway between tests using haplotypes and tests of single markers. One alternative [Zhang et al., 2003] builds on the idea of LD extending across multiple SNPs in the marker set. This approach involves smoothing the pattern of association in a target region, much as has been done in the context of fine-mapping [Lazzeroni, 1998; Cordell and Elston, 1999; Conti and Witte, 2003]. Then a smoothed curve is tested for a significant mode.

The other strategy is the Hotelling T^2 test and related procedures [Fan and Knapp, 2003; Chapman et al., 2003]. In the simplest case, in which only the allele counts are included in the design matrix, this approach is essentially equivalent to a particular multiple regression based test. We call the test statistic associated with this multiple regression approach T_R. This test simultaneously queries the relationship between the phenotype and the single-locus allele counts and their linear combinations. Clearly, this test can detect an association between the phenotype and the allele count for a particular locus. More subtle is its ability to detect an association between the phenotype and a linear combination of allele counts for a subset or all of the loci. The latter might occur if the liability allele were unmeasured but well predicted by a set of measured SNPs. Alternatively, to test for association between the phenotype and haplotypes, a more complex design matrix would be required.

Effective "single locus" tests require careful choice of tag SNPs to ensure all variants in the gene/genomic region are either tested directly or are tested indirectly via genotyping of SNPs in tight LD with them. Various methods for choosing tag SNPs from the larger set are available [Zhang et al., 2002b; Ackerman et al., 2003; Ke and Cardon, 2003; Meng et al., 2003; Sebastiani et al., 2003; Stram et al., 2003, Carlson et al., 2004]. We have recently introduced another method to choose tag SNPs, called H-clust, which uses hierarchical clustering [Rinaldo et al., 2005] to analyze data on genotypes, rather than haplotypes (Fig. 1).

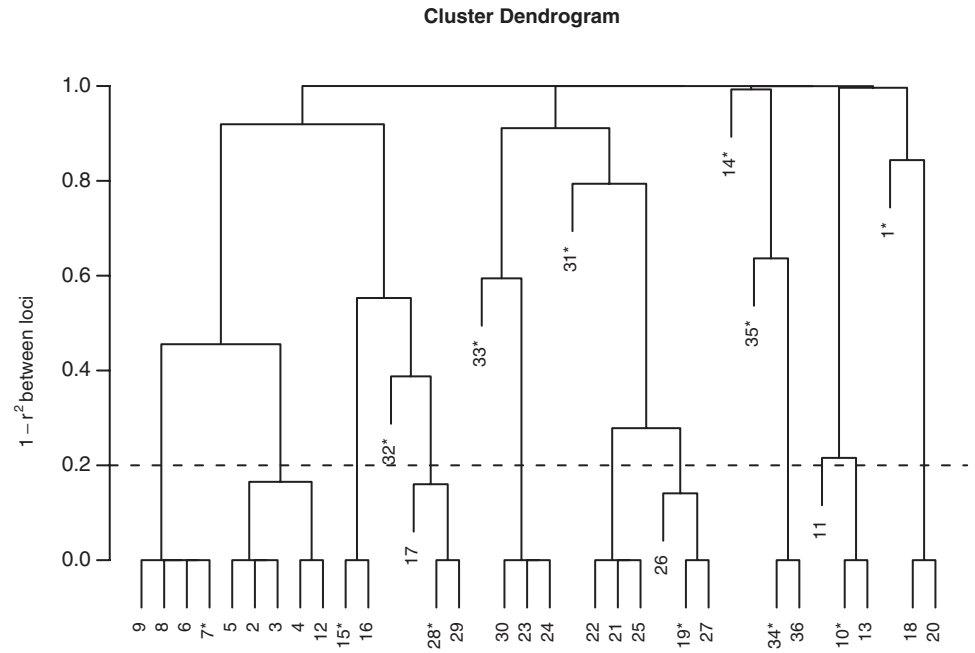

Fig. 1. Hierarchical clustering of SNPs in a gene illustrate potential block structure. The data are from individuals of European descent from the gene labeled JANK3 in the Seattle SNP database. The hierarchical clustering was performed using Pearson's squared correlation coefficient. Using a cutoff value of $\eta = 0.8$, 12 tag SNPs would be selected (e.g., starred SNPs).

The results of Chapman et al. [2003] suggest regression approaches, based on genotypes from tag SNPs, are usually more powerful than tests based directly on data on haplotypes. Because their regression approach can be conceptualized as a middle ground between haplotype-based tests and single locus tests, we wondered if the results of Chapman et al. [2003] could be taken even further. Specifically, following the logical progression, would single-locus tests prove generally more powerful than the regression-based tests? We answer this question by exploring various approaches to ''single-locus'' testing. In addition to the regression-based approach (T_R), we evaluate the smoothing approach to test for association. In their original work, Zhang et al. [2003] presented a Bayesian method, called BARS, to implement this integrated testing. We evaluate a simpler approach, which fits a nonparametric curve to the pattern of single-locus test statistics and then finds the maximum value of the fitted curve. We call the test associated with this procedure T_S. We contrast these two integrated procedures with two procedures based on a series of single locus tests: for T_B, the maximum test statistic is calibrated by using Bonferroni correction; for T_P, the maximum test statistic is calibrated by using permutation.

To evaluate the power of these tests, we assume a liability allele (LA) is relatively common and has little effect of increasing liability to disease. We consider models in which there is a single LA per gene and models with a *cis* interaction between pairs of causal variants. To evaluate power, we simulate multilocus genotype data with various levels and patterns of LD by using the evolutionary coalescent model. To determine the effect of LD patterns, we place the LA in blocks of high LD and outside of such blocks. We investigate the power to detect association by using either tag SNPs or the full set of ''common'' SNPs in the targeted region. We also compare the performance of these methods for tag SNPs selected at various levels of coverage.

METHODS

GENERATING DATA

To obtain data with an LD structure similar to that observed in human genes, we simulated ten ''genes''' with correlation patterns quite like those found in the SeattleSNP database. To generate haplotypes with blocky patterns of LD, we used a modification of Hudson's [2002] MS program [Wall and Pritchard, 2003] as described in Rinaldo et al. [2005]. The program used the coalescent model with a variable recombination rate over the gene to produce recombinational ''hot spots'' separated by ''cold regions.'' For each of the ten simulated genes, we modeled 2–4 cold regions, with lengths of these regions similar to those observed in particular genes in the SeattleSNP database. Each gene was 40–50 kb in total length. The scaled recombination rate over the entire region was r times the length of the region in base pairs, and r varied between 4 and 8×10^{-3}. The per basepair recombination rate varied so that it was 45 times greater in hot spots than in cold regions. The mutation rate was chosen to be $\theta = 5.6\times10^{-4}$/bp, which yields, on average, the number of common SNPs per kb typically observed in the SeattleSNP database. A sample of haplotypes was produced for each gene and the resulting common SNPs (those with minor allele frequency ≥ 0.05) were retained. For association analysis, the haplotypes were randomly paired to create individuals.

Certain portions of the simulated genes have far less haplotype diversity than expected if the alleles at multiple markers were independent, while other sets of adjacent markers behave almost independently. Regions with limited haplotype diversity are often called blocked. The location of apparent blocks was formally determined using the EB blocking methods described in Rinaldo et al. [2005]. The EB blocking method aims to find blocks using an approximate likelihood model, and a parsimony criterion to penalize for model complexity. Unlike other blocking methods, EB allows for adjacent, distinct blocks and also multiple, approximately independent SNPS separating blocks. The goal of the method is to partition the SNPs into blocked and unblocked regions. Specifically, the model uses a multinomial likelihood for blocked regions and an independent Bernoulli model for unblocked regions. The blocking structure is chosen to minimize the AIC criterion when a penalty is taken for the number of parameters in the model.

The disease model was chosen to mimic the type of data likely to be encountered when studying a complex disease. We simulated data for a case-control design, in which at least one LA with moderate effect was within a gene. A LA was selected from among SNPs sampled in the gene if it was either clearly in an apparent block or clearly outside of a block. The relative

frequency of the selected LA in the population of simulated haplotypes is denoted by p. The choice of LA varied by simulation condition with an aim to illustrate a diversity of conditions. Overall we simulated four conditions per gene: two with a LA inside a block and two with a LA outside a block.

To define a penetrance model for a liability locus, we first set the attributable fraction (AF) [Pfeiffer and Gail, 2003], which is defined to be one minus the ratio of the probability of affection, given that an individual has no copies of the detrimental allele, over the prevalence (K). With a little algebra, AF can be shown be the Levin's population attributable risk, and is thus easily related to the odds ratio and relative risk, two standard epidemiological quantities.

Let T be a test statistic based on the log odds ratio. For an additive genetic model, the power of a single test statistic T, evaluated for genotypes from the LA itself, can be expressed as a function of the sample size N, p, K, and AF. We fixed N at 500 cases and 500 controls, varied K and p, and solved for the value of AF that maintains power at 0.92, approximately. This trio of values is well approximated by a quadratic function, $AF = 0.05 - 0.07K + 0.57p - 0.15p^2 - 0.46Kp$. We then set $K = 0.01$. Now, for each choice of LA and associated p, we choose a corresponding AF to maintain a constant level of power at the LA. To generate the cases and controls, it remains to determine the penetrance function f_j, defined as the probability of affection, given j copies of the LA. For an additive model, $f_0 = (1 - AF)K$, $f_1 = f_0 + (AF\,K)/(2\,p)$, and $f_2 = 2f_1 - f_0$.

To generate cases and controls, haplotypes are drawn from a population and assigned as follows: (1) Select a LA and label one variant form as a '1.' Compute p in the population. (2) Calculate AF, f_0, f_1, and f_2. (3) Randomly select two of the haplotypes to form an individual. (4) Let x be the number of copies of the 1 allele at the liability locus. (5) Assign this individual case status if a randomly selected number between 0 and 1 is less than f_x. Otherwise assign it control status. (6) Repeat steps 3 to 5 until the required number of cases and controls are obtained.

To adapt this simulation scheme to accommodate a 2-SNP *cis*-model for liability, it suffices to define a "super-allele." Say a "1" occurs when both LAs are ones. Otherwise, the new allele is coded as a 0. Now proceed with the algorithm described above to generate cases and controls. The parameter p is the fraction of haplotypes that are of the inferred 1 type. LAs for the *cis* model were chosen in two ways: (1) both LAs were in the same block; or (2) one LA was in a block and the second was nearby, but not in the same block.

TESTING FOR ASSOCIATION

The simulated data are naturally summarized by a 2×2 table, with entries n_{ij} being the number of alleles of each type ($i = 0$ or 1) found in cases ($j = 1$) and controls ($j = 0$); for numerical stability add 1/2 to each value of n_{ij}. The test statistic T is defined as the absolute value of the log odds ratio, $|\log(n_{11}n_{00}) - \log(n_{10}n_{01})|$, divided by its standard deviation $(\sum_i \sum_j 1/n_{ij})^{1/2}$. The simplest multilocus test for L loci is $T_m = \max\{T_l, l = 1, \ldots, L\}$. If the rejection region is determined by using a Bonferroni correction, we call this test T_B. The next simplest test is also based on T_m, but with the rejection region determined based on permutations of the cases and controls (T_P).

We use T_S to test for a significant mode of a series of test statistics. Specifically, for a sample of L markers with physical locations, $b_1, b_2, \ldots, b_L$ and test statistics $T_1, T_2, \ldots, T_L$, we fit a smooth curve $g(\cdot)$ to the data that allows for additional noise, $e_1, e_2, \ldots, e_L$: $T_l = g(b_l) + e_l$. Let $S_l = \hat{g}(b_l)$ be the fitted value of the curve at location l, and let the test statistic be $T_S = \max\{S_1, S_2, \ldots, S_L\}$. The test is evaluated by repeated permutations of case-control status to produce a sequence of test statistics under the null hypothesis of no association. If T_S is greater than 95% of the statistics from the permutation-based series, it is declared significant.

T_S has several noteworthy properties: it will be powerful when there are strong modes in the data, but it does not force unimodality; it will have little power when the signal is very local, such as a single large T_l, because it will tend to smooth out the signal; and due to this smoothing, it will tend to be more robust than other statistics.

If no single SNP in the tag SNPs is highly correlated with the causal SNP, but together they predict it well, then a multivariate analysis is ideal. Chapman et al. [2003] pursued this avenue by using a score test. If the associated design matrix is based upon many single locus markers with no interactions, then the score test is essentially equivalent to the following generalized least squares regression approach. Let Y represent the phenotype (case vs. control) and let $\mu = E[Y]$. Using a logit link, model $g(\mu)$ as a linear function of a design matrix $\mathbf{X}$: $g(\mu) = \mathbf{X}\beta$. Assuming an

additive genetic model, create the design matrix from the allele counts (0,1,2) at each SNP. A test for $\beta = 0$ simultaneously tests whether any particular multilocus allele count, or linear combinations of these, is associated with the phenotype. This is the T_R test.

SELECTING TAG SNPs

Like that observed for human genes, the simulated data contained a large number of correlated SNPs. To analyze only a subset of those SNPs, namely the tag SNPs, we used the H-clust method [Rinaldo et al., 2005] to select tag SNPs. To "tag" a genomic region, we must first decide how tightly associated the genotyped and unmeasured SNPs (those not genotyped) must be.

H-clust is a simple method that can be used to rapidly identify a set of tag SNPs based upon genotype data [Rinaldo et al., 2005]. H-clust does not require haplotype estimation. Instead, H-clust recodes the genotypes as 0, 1, 2, to denote the number of copies of a particular allele. It then computes a similarity matrix based upon the square of Pearson's correlation (r_p^2) between allele counts. H-clust consists of two stages: hierarchical clustering to group correlated SNPs based on the similarity matrix, and choosing a SNP or SNPs to represent clusters. H-clust chooses a set of tag SNPs such that no unmeasured SNP has $r_p^2 < \eta$ for at least one tag SNP.

Hierarchical clustering can be represented as a dendogram in which any two SNPs diverge at a height equal to the difference between one and the squared correlation between the two observations. The clusters are obtained by declaring SNPs to be in the same cluster when they are found in the same sub-tree below a preset value, say $1 - \eta$. The clustering method yields SNPs that are in multiple SNP clusters and SNPs that form clusters of size one. H-clust uses the "complete linkage" method.

The next step chooses the representative tag SNP for each cluster. The method chooses the SNP that is most correlated with all the other SNPs in the block. If multiple SNPs show equal correlation, then the one in the middle is chosen as the tag SNP.

A stepwise regression procedure can be used to obtain an even more parsimonious set of tag SNPs. After selecting a tag SNP set using H-clust, subsets of potential tag SNPs are assessed based upon their ability jointly to predict the allele counts of unmeasured SNPs (on average). For each unmeasured SNP, the coefficient of determination, or R^2, for predicting the allele count in the region with a given subset of SNPs is measured. We use a backward elimination procedure to determine which SNPs to drop. With this sequential procedure SNPs are dropped from the tag SNP set sequentially until the average R^2 drops below the prespecified level called η^*. We call this method H-clust stepwise.

RESULTS

SIMULATED GENES

By using the coalescent model, we simulate data from ten "genes" (A–J). Features of these genes roughly match those seen in the SeattleSNP database (see Electronic-Database Information) [Crawford et al., 2004], and they also reflect the inter-gene variability seen in that database. Thus, although the genes were simulated by using similar evolutionary parameters, they vary in the number of common SNPs from 44 to 99 and the apparent blockiness (Table I). In addition, haplotype block structure is quite apparent for some simulated genes, and less so for others (results not shown). The strength of the correlation within apparent blocks is summarized by computing the average pairwise correlation within apparent blocks and ranges from 0.46 to 0.77 (Table I).

TAG SNPs

The number of tag SNPS required to cover a gene also varies across genes. For the simulated genes, and an H-clust r_p^2 cutoff value of $\eta = 0.8$, the modal number of tag SNPs varies considerably,

TABLE I. Features of 10 simulated genes[a].

			Number of Tag SNPs			
Gene	Total	Mean correlations	0.8	0.6	0.4	0.2
A	44	0.46	18	14	12	11
B	46	0.61	21	18	13	10
C	65	0.62	18	15	13	10
D	52	0.62	17	16	13	9
E	85	0.62	27	19	14	11
F	49	0.67	24	17	15	12
G	49	0.70	21	16	14	9
H	58	0.74	12	9	7	5
I	99	0.75	25	20	16	11
J	56	0.77	19	15	11	9

[a]The number of common SNPs total (minor allele frequency ≥ 0.05) and the modal number of tag SNPs for four levels of stringency (0.8, 0.6, 0.4, 0.2). The mean correlation is the average pairwise correlation among SNPs within blocks.

and depends more upon the fraction of SNPs in tight LD than upon haplotype blocks per se. Lowering η from 0.8 to 0.2 reduces the modal number of tag SNPs, for example, from 12 to 5 for gene H. For some genes, however, the tag SNP set is substantial even at $\eta = 0.2$ (Table I) because some SNPs are nearly uncorrelated with all others (Fig. 1).

To assess coverage provided by the H-clust tag SNPs, we computed the coefficient of determination, R^2, for predicting a SNP genotype using only the genotypes of the tag SNPs [Stram et al., 2003; Chapman et al., 2003]. We computed this quantity for every common SNP in a gene and for each level of η (Table II). Overall, the average R^2 is quite high (≥ 0.81) even when $\eta = 0.2$. The minimum R^2 shows more variability but it is also fairly high, suggesting that H-clust chooses tag SNPs that provide good coverage of a gene.

Within a cluster of highly correlated SNPs, some of which are perfectly redundant, the choice of a tag SNP can be arbitrary. Hence, numerous sets of tag SNP with minimal overlap could perform almost identically in an association study. By comparing the R^2 of the tag SNPs selected by H-clust to a validated method of selection developed by Carlson et al. [2004], Rinaldo et al. [2005] show the tag SNP sets to be almost identical in predictive power regardless of the cutoff values ($\eta = 0.8, 0.6, 0.4$). On the other hand, H-clust sometimes selects slightly fewer SNPs.

When stepwise-H-clust was used to select the tag SNPs, a much more parsimonious collection of SNPs was identified. For an R^2 of 0.8, only a modal value of 5 SNPs were required for each gene; for an R^2 of 0.5, only 2 SNPs were required (except J, which only required 1).

TABLE II. Coefficient of determination obtained by using Tag SNPs to predict common SNPs in the ten simulated genes[a]

	Min R^2				
Gene	0.8	0.6	0.4	0.2	Avg R^2 (0.2)
A	0.942	0.786	0.789	0.634	0.951
B	0.959	0.914	0.724	0.671	0.947
C	0.919	0.835	0.716	0.705	0.947
D	0.936	0.899	0.721	0.490	0.924
E	0.968	0.935	0.703	0.558	0.909
F	0.929	0.836	0.726	0.325	0.883
G	0.950	0.765	0.741	0.512	0.893
H	0.894	0.830	0.783	0.451	0.812
I	0.928	0.791	0.791	0.456	0.948
J	0.886	0.894	0.645	0.565	0.903

[a]Minimum R^2 is given for four levels of stringency $\eta = (0.8, 0.6, 0.4, 0.2)$ for selecting tag SNPs. The average is given for one $\eta = (0.2)$.

PATTERN OF TEST STATISTICS

We tailored the parameters for the disease model to the frequency of the alleles at the locus chosen to confer liability so that there would be roughly 92% power to detect its effect on liability. The realized differences in power across the ten genes were then due to the local patterns of LD within the genes, how the LA fits into that pattern of LD, and the choice of tag SNPs.

To make the simulations more concrete, we present some summary statistics for the presence of the disease locus in the set of tag SNPs and pattern of test statistics. Of the 40 simulation conditions (10 genes, 4 disease SNPs taken one at a time), there were 250 simulations. Averaged over these 10,000 simulations, the disease variant was present in the tag SNP set 30% of the time when $\eta = 0.8$ and 21.2% when $\eta = 0.2$. Its distribution of presence/absence was largely bimodal, with some disease variants always present in the tag set and others always absent.

For two choices of tag SNPs ($\eta = 0.8$ and 0.2), we explored the pattern of test statistics, $T_1, \ldots, T_m$, by viewing them as single tests for association. *For descriptive purposes only*, we call a test significant if it has a p value less than 0.05. By this definition we found, on average, 32.5% (34.5%) of the tag SNP markers to be significant for $\eta = 0.8$ ($\eta = 0.2$). Furthermore, the percentage of significant tests is typically greater when the LA is located in a block (37.1% for $\eta = 0.8$; 39.5% for $\eta = 0.2$) than out of a block (27.8% for $\eta = 0.8$; 29.6% for $\eta = 0.2$). For some scenarios, the LA was never included in the tag SNP set. Stratifying to the scenarios for which the LA was included in the tag SNPs, we found that the LA was significant 92.8% (90.3%) of the time for $\eta = 0.8$ ($\eta = 0.2$). Furthermore, we found that the LA was the maximum statistic 62.3% (75.6%) of the time for $\eta = 0.8$ ($\eta = 0.2$). This percentage differed substantially by whether the LA was in or out of a block. For $\eta = 0.8$, the LA yielded the maximum statistic 75.3% of the time when it was out of a block, compared to only 49.2% of the time when it was in a block. When the LA was out of a block, it achieved significance 91.9% of the time for $\eta = 0.8$ and 86% for $\eta = 0.2$. Alternatively, when the LA was in a block, it achieved significance 93.6% of the time for $\eta = 0.8$ and 94.6% for $\eta = 0.2$. The

power of the individual test statistics (T_i) depends on the allele frequencies. A tag SNP that is in LD with the LA can often yield a larger test statistic than the LA itself.

SUMMARY OF POWER

Our initial comparison of the competing methods assumes that all common SNPs are genotyped. Under this scenario, the T_R approach is excluded because it targets selection of a subset of SNPs. We recognize that genotyping all SNPs in a gene is not common practice. On the other hand, such a scenario is conceivable if, for example, some forms of pooled genotyping are performed (e.g., pooled sequencing). Moreover, the comparisons are quite instructive (Fig. 2). For every gene, T_P is substantially more powerful than T_B because the former automatically accounts for the correlation among SNPs. Averaging over all conditions, T_P is significantly more powerful than T_S. On a per gene basis, however, T_S is typically more powerful than T_P provided the LA is in a block. The pattern reverses when the LA is out of a block.

Selection of tag SNPs by H-clust. When only tag SNPs are genotyped, T_P is again more powerful than T_B, but the advantage is smaller because the tag SNPs are less correlated. For $\eta = 0.8$, T_P has 7% more power than T_S, on average, but when the LA is within a block, the approaches are equal in average power. Indeed, for some conditions T_S is substantially more powerful than T_P (Fig. 3, top left). Finally, on average, T_P shows a 21% advantage in power over T_R at $\eta = 0.8$ (Fig. 3, top). On a per gene basis, T_R is sometimes nearly equal to T_P, but never substantially better than it.

For $\eta = 0.2$, the power of T_R increases by an average of 8%, which can be attributed to two features: the decrease in degrees of freedom for the test and the relatively high predictive ability of the smaller set of tag SNPs. While we would predict that the power of T_P should decrease as η decreases, T_P is surprisingly resilient to changes in the number of SNPs genotyped. The average drop

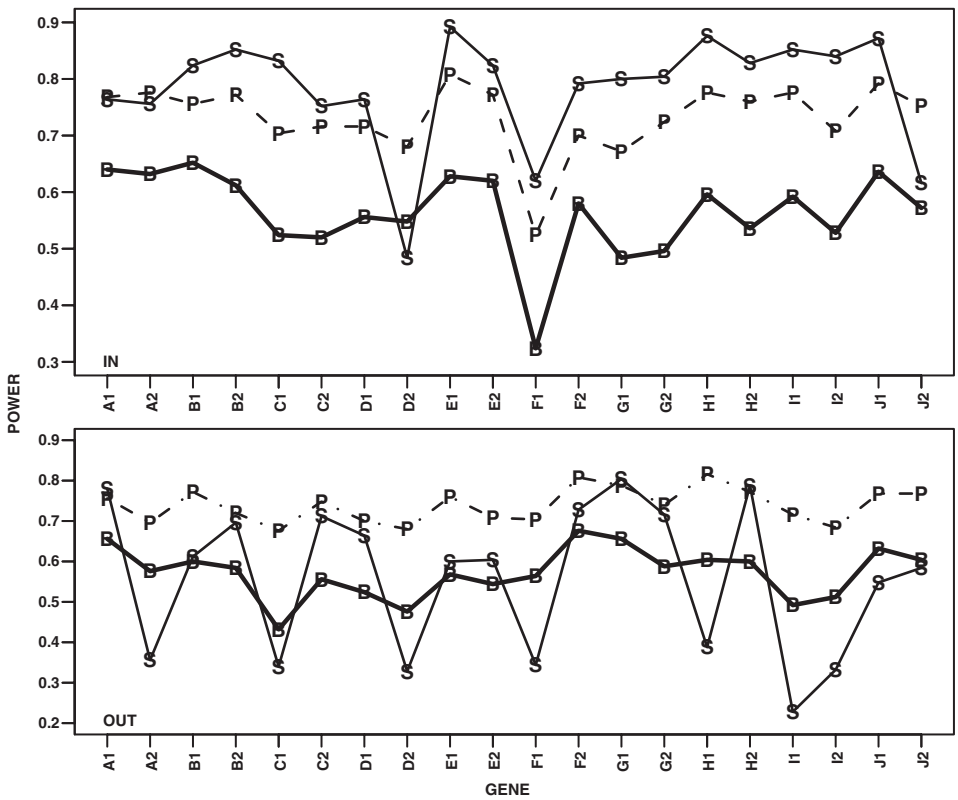

Fig. 2. Power of competing tests of association applied to ten simulated genes by using all common SNPs. The methods are B=T_B, P=T_P, and S=T_S. Two LAs are selected under each of two conditions for genes labeled A–J. The LA is either located in a block (top) or out of a block (bottom).

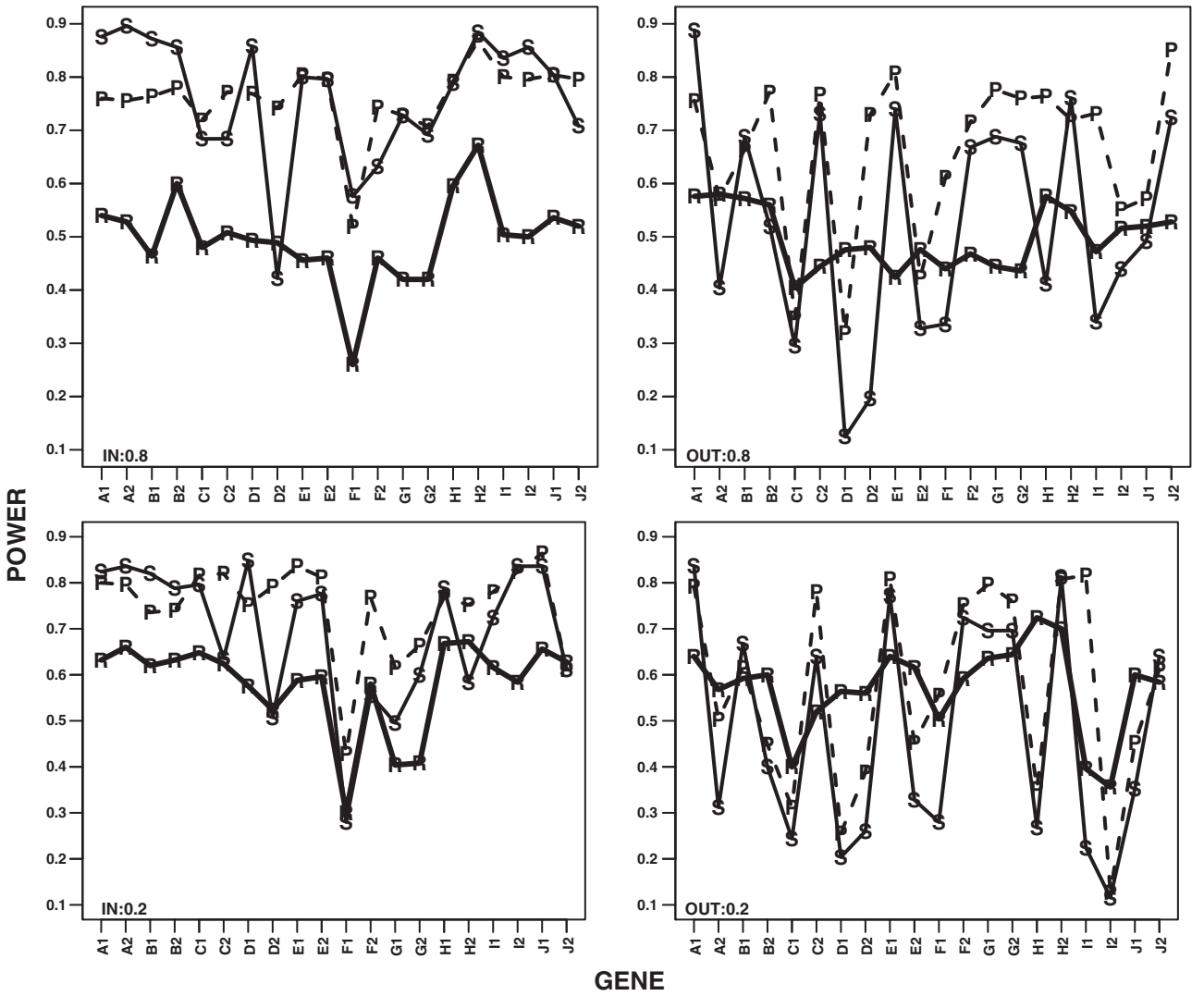

Fig. 3. Power of competing tests of association applied to ten simulated genes by using selected tag SNPs. The methods are R=T_R, P=T_P, and S=T_S. Two LAs are selected under each of two conditions for genes A–J. The LA is either located in a block (top) or out of a block (bottom). Left: H-clust cutoff of $\eta = 0.8$. Right: H-clust cutoff of $\eta = 0.2$.

in T_P's power is 5%. Consequently, T_P is more powerful than T_R even when the number of tag SNPs is relatively small. Still, it is worth noting that the powers of the two tests are roughly equal when the LA is out of a block and $\eta = 0.2$ (Fig. 3, bottom right).

To assess the impact of the exclusion of the LA from the tag SNP set, we drop the LA from the tag SNP set and reassess R^2. As measured by R^2 and for $\eta = 0.8$, the LA for most genes was adequately predicted by the other SNPs in the tag set, and consequently power to detect the LA is essentially unchanged (results not shown). Two exceptions occurred when the LAs were out of a block (genes B and J). In these two instances T_P and T_S were substantially reduced (Table III), while T_R was less

TABLE III. Decrease in power for three tests of association, T_P, T_S, and T_R, when the LA is removed from the Tag SNP set[a]

Gene	Position	T_P	T_S	T_R
B	In	0.00	0.01	−0.08
B	In	−0.04	−0.01	0.06
B	Out	0.08	0.08	0.05
B	Out	0.30	0.17	0.10
J	In	0.01	0.06	0.04
J	In	0.12	0.08	−0.00
J	Out	0.05	0.03	0.01
J	Out	0.26	0.24	0.11

[a]Tag SNPs were selected using H-clust ($\eta = .8$) and then the LA was removed from the tag SNP set, if it was originally in it. The LA was either located in a block (In) or out of a block (Out). Results are shown for only two of the ten genes. Entries are the difference in power before minus after removing LA.

affected. Nevertheless, the average power of T_P still exceeded that of T_R. Aside from these exceptions, the power was essentially unchanged when the LA was dropped from the tag SNP set. At $\eta = 0.2$, the tag SNPs happened to not include the LA, and hence dropping the LA would have no effect on the average power.

Selection of tag SNPs by H-clust stepwise. Our results thus far show that the power of T_R increases and the power of T_P decreases as the number of tag SNPs representing the gene decreases. Therefore, we continue to compare these two statistics when tag SNPs are selected by H-clust stepwise. As demonstrated in a previous section, when H-clust stepwise is applied to the data from these simulated genes, roughly 5 and 2 tag SNPs are selected for $\eta^* = 0.8$ and $\eta^* = 0.5$ respectively. For these levels of η^*, T_R is often slightly, but not significantly, more powerful than T_P, on average. On a per gene basis, the methods can differ substantially in their power (Fig. 4).

Interaction. When two LAs are present in each gene with a *cis* interaction, the results are surprisingly similar to those obtained for a single LA (Fig. 5). Tag SNPs are selected by setting $\eta = 0.8$. T_P performs best overall, with power 17% better than T_R. The power of T_S is essentially equal

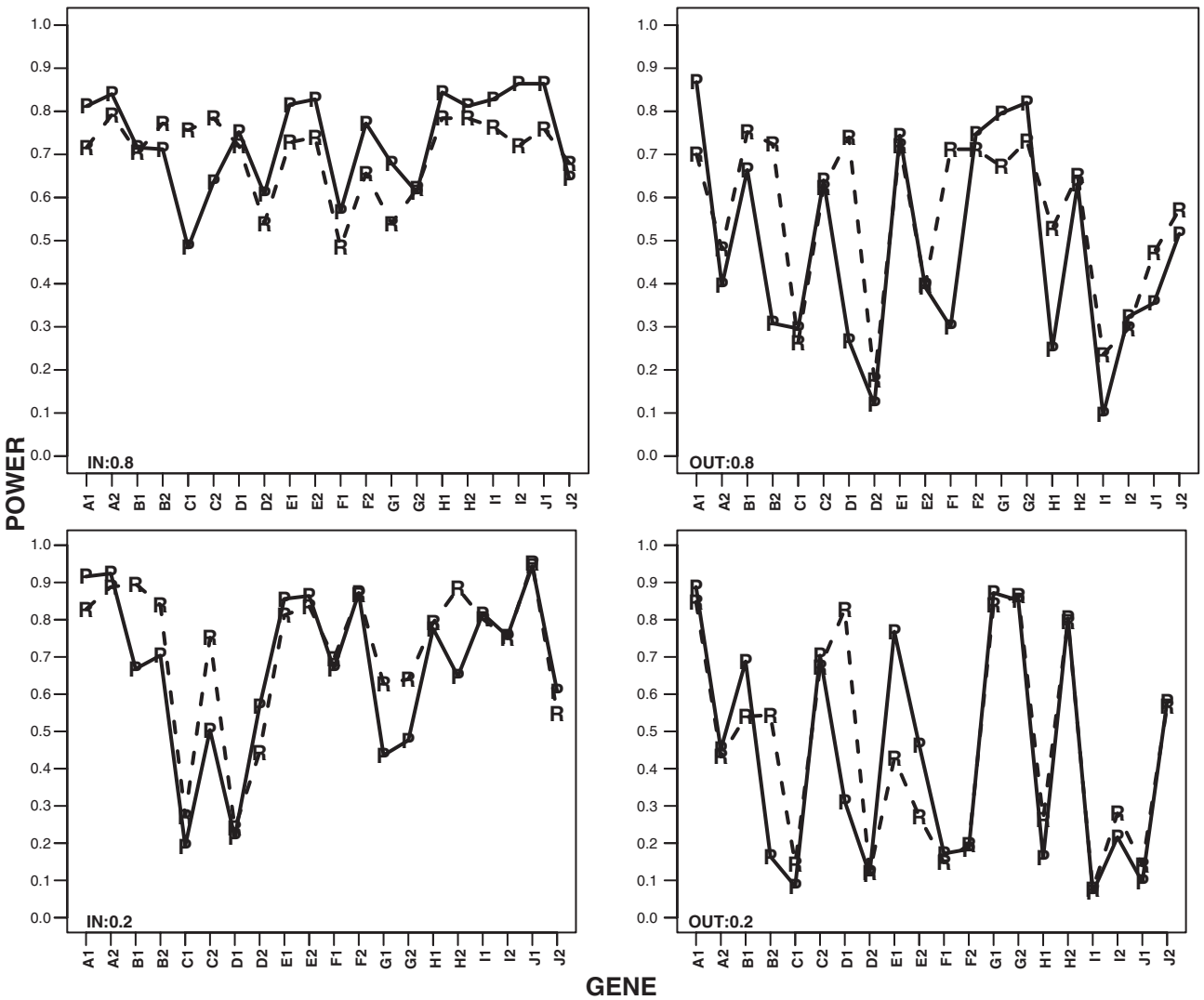

Fig. 4. Power of competing tests of association applied to ten simulated genes by using selected tag SNPs. The methods are R=T_R and P=T_P. Two LAs are selected under each of two conditions for genes A–J. The LA is either located in a block (top) or out of a block (bottom). Left: H-clust stepwise cutoff of $\eta^* = 0.8$. Right: H-clust stepwise cutoff of $\eta^* = 0.5$.

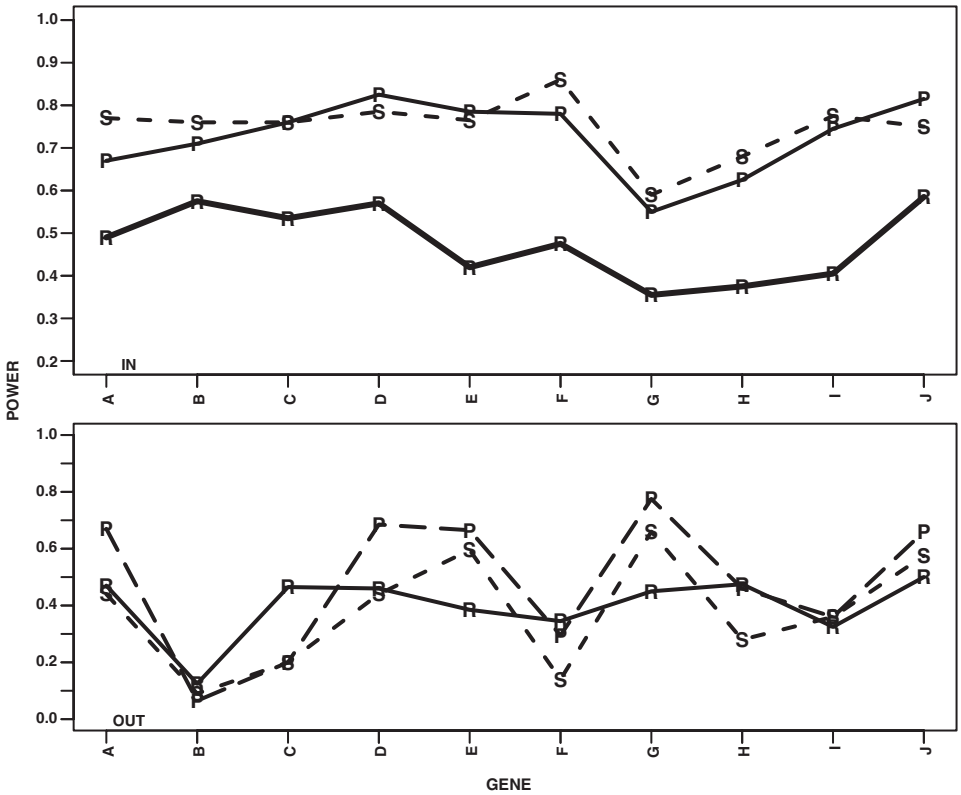

Fig. 5. Power of competing tests of association applied to ten simulated genes by using selected tag SNPs and a *cis* model for LAs. The methods are R=T_R, P=T_P, and S=T_S. Two pairs of LAs are selected under each of two conditions for genes A-J. One LA from each pair is selected inside a block; the second LA is either located in a block (top) or out of a block (bottom). The H-clust cutoff is $\eta = 0.8$.

to T_P when both LAs are drawn from the same block. Otherwise, the power of T_S is 11% lower than that of T_P. For the cis model, the most notable power result occurs when one LA is in a block and the other is out (Fig. 5, bottom). In this scenario, for each testing procedure, the power is considerably reduced compared to the single LA scenarios investigated previously.

DISCUSSION

In this research, we investigate four procedures for testing for association between variation in a human gene and complex disease: the maximum of a series of test statistics corrected for multiple testing by the Bonferroni procedure, T_B, or by permutation of case-control status, T_P; a procedure that tests the maximum of a smoothed curve fitted to the series of test statistics, T_S; and the Hotelling T^2 procedure suggested by Fan and Knapp [2003] and Chapman et al.[2003], T_R. The results of Chapman et al. [2003] suggest that T_R is generally more powerful than omnibus tests of haplotype/disease association.

We test these procedures by using simulated data that have many of the features of data from human populations. Specifically, by using a program to mimic human evolution, we simulate genetic variation with correlation structure similar to that observed in genes from the SeattleSNP database. Within this context, we varied the location of liability alleles (LA) within genes, placing LAs in regions with extensive local LD and regions with very little LD.

Although results varied depending upon the correlation structure of SNPs within a gene, the density of tag SNPs, and the placement of the LA, we observed a fairly consistent pattern in the performance of association tests. Overall, T_P performed at least as well and usually better than the other test statistics. Permutation of case-control status, as is done for T_P, is a simple means of developing a null distribution for test statistics while also accounting for the LD pattern among SNPs in the gene or genomic region. For this reason, T_P typically had much better power than T_B, the test adjusted by Bonferroni correction. On the other hand, in some cases T_B is not substantially underpowered relative to T_P, especially

when tag SNPs have been selected and thus the correlation among SNPs is not large.

When a dense set of SNPs is genotyped and the LA is located within a highly correlated set of SNPs (i.e., a haplotype block), T_S is the most powerful approach. This advantage occurs because several SNPs, in addition to the LA, are associated with the phenotype. Hence, the pattern of the signal for association is quite reliable. T_S exploits the redundant information carried by the other SNPs in the block, thereby greatly reducing the variance of the test statistic and hence increasing the power of the test. Although T_P accounts for the LD between markers when computing the cutoff value for rejection, it does not utilize the consistent pattern of association. Hence, it fails to assimilate supporting evidence for a signal. If the LA is not in LD with alleles at neighboring markers, however, T_S is unlikely to detect a clear signal of association. Likewise, when only a small, selected set of tag SNPs is genotyped, T_S loses its power advantage over T_P. The fitted curve inherent to T_S is not likely to be influenced substantially by a single high point in the set of markers. This feature makes T_S robust to false positives, but robust statistical procedures come at the price of lost power in some settings.

Perhaps the most surprising result of our investigation is the 21% advantage in power of T_P over T_R when tag SNPs are selected by H-clust and $\eta = 0.8$ (Fig. 3). To understand this result, first consider the theoretical principles underlying these two tests. In the context of multiple testing, T_P protects against m single-locus tests for association (with an implicit adjustment for the correlation among these tests). Alternatively, the T_R test is equivalent to simultaneously testing for association with an effect due to any linear combination of the tag SNPs. Consequently, this test is able to detect a signal when the LA is not measured and not highly correlated with any of the tag SNPs, provided it is well predicted by a linear combination of the tag SNPs. This would seem to offer the T_R approach a substantial advantage, but there is a price for this flexibility. The set of all linear combinations is a large set of hypotheses to consider simultaneously. There is a substantial penalty in power due to an implicit penalty for multiple comparisons. This penalty is especially severe when many tag SNPs are used.

On the other hand, when a much sparser set of tag SNPs is selected by using H-clust stepwise and $\eta^* = 0.8$ or $\eta^* = 0.5$, the average performance of T_P and T_R is virtually indistinguishable. The average performance conceals some distinct differences in power for the methods present for particular SNPs in particular genes. These differences occur because some unmeasured SNPs can only be predicted by a set of tag SNPs, and are not well represented by any single tag SNP. If such a SNP is the LA, then T_P will perform poorly relative to T_R. Because the LA in a gene is unknown, it would be impossible to predict, a priori, which method will be more powerful for a sparse set of tag SNPs. It is important to note, however, the decrease in average power for the scenarios in which a sparse set or denser set of tag SNPs is selected (contrast Fig. 4 and Fig. 3). For example, for T_P and $\eta = 0.8$, the average power is 71%; when $\eta^* = 0.8$, the average power drops to 61%. For the same comparisons and T_R, the average power is 49.6 and 63.1%. As is typical, the investigator faces a trade-off between cost and power.

The more subtle point is that two methods under investigation gain power as the number of SNPs increases (T_S, T_P), while another loses power (T_R). Therefore, to thoroughly investigate a gene, a two-stage analysis might be appropriate. For tag SNPs selected using H-clust stepwise, conduct an initial analysis using T_P and T_R. Then, if there are no significant findings, genotype a larger set of tag SNPs, which would be selected by H-clust. In this way, one obtains good power to obtain a signal while still minimizing the amount of genotyping and minimizing the risk of false negatives.

Alternatively, if all genotyping were to be performed simultaneously, one could envision a two-stage testing strategy: first test using T_R and a smaller set of tag SNPs; and second test using T_P and a larger set of tag SNPs. It is important that any testing via T_R be done using a minimal panel of SNPs, such as those selected by H-clust stepwise. Correlations among tag SNPs do not have a detrimental effect on T_P and T_S. Indeed, it could enhance the power of these tests.

In our investigations, we assume that the LA was relatively common (minor allele frequency greater than 5%) and that an accurate physical map of the common SNPs is available from which to select a set of tag SNPs. In this situation, Chapman et al. [2003] argued that the T_R approach is more powerful than a full haplotype approach. We take their logic a step further to conclude that single locus tests are generally more powerful than the T_R approach, except when very few tag SNPs are tested.

Our investigations also assume a complete set of common SNPs is available. If only a small subset is available, there could be crucial gaps in the coverage of the region. When there are gaps, procedures such as H-clust cannot yield an optimal tag SNP set. Moreover, the relative power of T_P and T_R is unpredictable, even on average, because some unmeasured SNPs have a very low pairwise correlation with each of the tag SNPs. In this setting, we would not be surprised to find that T_R can be slightly more powerful, on average, than T_P.

Our results and those of Chapman et al. [2003] should not be over-interpreted. These results do not suggest that knowledge of haplotype structure is of no help in the analyses of genetic association. The multivariate correlation structure of the SNPs determines the choice of tag SNPs. Tag SNPs cannot be selected appropriately without a careful preliminary analysis of the full set of SNPs. At least partial information to select tag SNPs is emerging from the HapMap project [Gibbs et al., 2003]. And accessory information, such as the evolutionary relationships among haplotypes [Seltman et al., 2003], should also prove valuable for association analysis.

Our exploration of the pattern of single locus test statistics revealed that it is hard to determine which particular SNP under investigation affects the phenotype, if indeed any one has a causal effect. Often, several SNPs are significantly associated with the phenotype and this is true whether or not the LA was measured. Although the problem is more challenging when the LA is located within a block of high SNP-to-SNP correlation, it exists outside of such blocks as well. It is reassuring to see that if the LA were to be included in the tag SNPs, it is often significantly associated with the phenotype (when no penalty for multiple testing is imposed); however, it does not always have the maximum test statistic. Clearly, it will be impossible to ascertain causality based on the significance of a series of single locus tests for association or even haplotypes [Seltman et al., 2001].

ELECTRONIC-DATABASE INFORMATION

SeattleSNPs Database, NHLBI Program for Genomic Applications, UW-FHCRC, Seattle, WA, http://pga.gs.washington.edu/. Results noted by ''results not shown'', currently in the ''Supplementary material'', will be posted at http://www.interscience. wiley.com/jpages/0741–0395/suppmat.

REFERENCES

Abecasis GR, Noguchi E, Heinzmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WO. 2001. Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191–197.

Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, Taylor T, Ward R, Molyneux M, Pinder M, Kwiatkowski DP. 2003. Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Biol 4:R24.

Carlson CS, Eberle, AM, Rieder JM, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120.

Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31.

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612.

Conti DV, Witte JS. 2003. Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. Am J Hum Genet 72:351–363.

Cordell, HJ, Elston, RC. 1999. Fieller's theorem and linkage disequilibrium mapping. Genet Epidemiol 17:237–252.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA. 2004. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610–622.

Daly MJ, Rioux JD, Schaffner SF, Hudson TH, Lander ES. 2001. High resolution haplotype structure in the human genome. Nat Genet 29:229–232.

Fan R, Knapp M. 2003. Genome association studies of complex diseases by case-control designs. Am J Hum Genet 72:850–868.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'and LY, et al. 2003. The International HapMap Project. Nature 426:789–796.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Ke X, Cardon LR. 2003. Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287–288.

Lazzeroni, LC. 1998. Linkage disequilibrium and gene mapping: an empirical least-squares approach. Am J Hum Genet 62:159–170.

Maniatis, N, Collins, A, Xu, CF, McCarthy, LC, Hewett, DR, Tapper, W, Ennis, S, Ke, X, Morton, NE. 2002. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc. Natl Acad Sci USA 99:2228–2233.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. 2004. The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584.

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG. 2003. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet 73:115–130.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Pfeiffer RM, Gail MH. 2003. Sample size calculations for population- and family-based case-control association studies on marker genotypes. Genet Epidemiol 25:136–148.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Rinaldo A, Bacanu S, Devlin B, Sonpar V, Wasserman L, Roeder K. 2005. Characterization of multilocus linkage disequilibrium. Genet Epidemiol 28:193–206 (this issue).

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF. 2003. Minimal haplotype tagging. Proc Natl Acad Sci USA 100:9900–9905.

Seltman H, Roeder K, Devlin B. 2001. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. Am J Hum Genet 68:1250–1263.

Seltman H, Roeder K, Devlin B. 2003. Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58.

Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. 2003. 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 55:27–36.

Wall JD, Pritchard JK. 2003. Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502–516.

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

Zhang K, Deng M, Chen T, Waterman MS, Sun F. 2002b. A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339.

Zhang X, Roeder K, Wallstrom G, Devlin B. 2003. Integration of association statistics over genomic regions using Bayesian adaptive regression splines. Hum Genom 1:20–29.