Genetic Epidemiology 33: 16–26 (2009)

Association Mapping by Generalized Linear Regression With Density-Based Haplotype Clustering

Robert P. Igo Jr,[1] Jing Li[2] and Katrina A.B. Goddard[1]*

[1]*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio*
[2]*Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, Ohio*

Haplotypes of closely linked single-nucleotide polymorphisms (SNPs) potentially offer greater power than individual SNPs to detect association between genetic variants and disease. We present a novel approach for association mapping in which density-based clustering of haplotypes reduces the dimensionality of the general linear model (GLM)-based score test of association implemented in the HaploStats software (Schaid et al. [2002] Am. J. Hum. Genet. 70:425–434). A flexible haplotype similarity score, a generalization of previously used measures, forms the basis for grouping haplotypes of probable recent common ancestry. All haplotypes within a cluster are assigned the same regression coefficient within the GLM, and evidence for association is assessed with a score statistic. The approach is applicable to both binary and continuous trait data, and does not require prior phase information. Results of simulation studies demonstrated that clustering enhanced the power of the score test to detect association, under a variety of conditions, while preserving valid Type-I error. Improvement in performance was most dramatic in the presence of extreme haplotype diversity, while a slight improvement was observed even at low diversity. Our method also offers, for binary traits, a slight advantage in power over a similar approach based on an evolutionary model (Tzeng et al. [2006] Am. J. Hum. Genet. 78:231–242). *Genet. Epidemiol*. 33:16–26, 2009.

Key words: fine mapping; linkage disequilibrium; high-density SNP genotypes

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 number: HL07567; Contract grant sponsor: NIH/NLM; Contract grant number: 008911; Contract grant sponsor: Case Western Reserve University; Contract grant sponsor: National Center for Research Resources; Contract grant number: RR03655; Contract grant sponsor: NIH/NIDDK; Contract grant number: P30 DK027651.
*Correspondence to: Katrina A.B. Goddard, Center for Health Research, Kaiser Permanente Northwest, 3800 North Interstate Avenue, Portland, OR 97227. E-mail: Katrina.AB.Goddard@kpchr.org
Received 16 February 2008; Accepted 7 May 2008
Published online 16 June 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20352

INTRODUCTION

With the advent of high-throughput genotyping methods for single-nucleotide polymorphisms (SNPs), it has become feasible to carry out association mapping studies that incorporate hundreds of thousands of genetic markers. Large sample sizes are necessary to identify disease-susceptibility loci with small effect. Population-based mapping studies rely on the correlation, or linkage disequilibrium (LD), between genetic variants influencing the trait and one or more markers being scanned. However, methods that examine SNPs in isolation fail to take advantage of the genetic information available in a dense marker map where significant LD exists among neighboring markers.

To incorporate multiple correlated markers, analyses can group alleles into haplotypes to improve both the accuracy of the genetic model and the power to detect association. Use of haplotypes will reduce the complexity of the model, while preserving the relevant interactions among markers. In addition, haplotypes best approximate the unit of expression of genetic variants, since single DNA molecules are transcribed [Akey et al., 2001]. Haplotype-based tests of association [e.g., Akey et al., 2001; Douglas et al., 2001; Fallin and Schork, 2000] can offer better power than genotype-based tests, especially when LD is high and a small number of haplotypes carry the causal variant [Thomas et al., 2003].

Grouping, or clustering, similar haplotypes to reflect their evolutionary history should improve our chances of locating a common causal variant. While haplotypes are susceptible to the evolutionary forces of mutation, recombination, and gene conversion, haplotypes carrying identical copies of a susceptibility locus should be more similar to each other than to haplotypes that do not carry the locus, on account of shared ancestry. In regions of high sequence diversity, large numbers of haplotypes reduce power by increasing the number of degrees of freedom (d.f.) of statistical tests. It is also hard to estimate haplotype effects in these diverse sequences because the sample size is very small for each haplotype. Through clustering, we can simplify the regression model while retaining most of the information relevant to disease-locus inheritance.

A number of statistical methods have been proposed in recent years to integrate haplotype grouping via similarity and/or shared ancestry in association mapping on population-based samples (see Table I). Coalescent-based

approaches attempt to reconstruct the population history of disease-causing mutations, under the assumption that they can be traced back to a single common ancestor [Kingman, 1982]. Several variations on the coalescent model have been implemented for association analysis [Liu et al., 2001; McPeek and Strahs, 1999; Morris et al., 2002, 2004; Zöllner and Pritchard, 2005]. Ultimately, none of the approaches rooted in explicit coalescent modeling allows for large-scale association analyses covering large portions of the genome.

Coalescent-based models are most applicable in the absence of recombination and selective pressure, neither of which is likely to apply in real case-control samples [Molitor et al., 2005]. Coalescent approaches are also extremely complex, and do not model attributes such as natural selection, ascertainment, and population substructure [Morris, 2005]. An alternative approach is to group haplotypes according to some measure of similarity that approximates the coalescent model indirectly. Thus, the goal is to simplify the model while retaining as much of the information as possible about the natural history of the sample.

One method that reduces the complexity of the inheritance model is constructing a cladogram of haplotypes based on their similarity, without attempting to reconstruct their full population history. The origin of these techniques is the cladistic analysis of Templeton et al. [1987], which employed an unrooted tree. Cladograms may also be constructed from a haplotype network [Seltman et al., 2003], in which haplotypes conferring similar disease risk are clustered based on inferred age. In contrast, Durrant et al. [2004] hierarchically arranged haplotypes by means of a distance metric based on the longest interval of consecutive matching alleles [the ''length measure'': Tzeng et al., 2003; or ''maximum identity length'': Bourgain et al., 2000], weighted to favor rarer alleles, which are more likely to share recent common ancestry.

Tzeng [2005] modified the cladistic approach to consider all possible relationships within a haplotype network. Their method allows haplotypes to be assigned to more than one cluster and weights both distance and haplotype frequency. A certain number of the most common haplotypes are selected as ''core'' haplotypes (assuming that more common haplotypes are more ancient). The remaining haplotypes are each assigned to one or more clusters using a probability score. The number of core haplotypes is selected based on the Shannon [1948] information criterion. Clustering is performed independently of the trait status and is geared to reducing the d.f. of statistical tests. The authors initially used the approach to improve power of Pearson's χ^2 test for association of haplotypes with binary traits. A later study [Tzeng et al., 2006] applied the clustering approach to the score test for haplotype association built on the general linear model (GLM) framework [Schaid et al., 2002], which improved the power of the score test on both dichotomous and continuous traits.

Cladistic techniques offer vast savings in computational intensity over the full coalescent-based approach. As a result, it is feasible to perform whole-genome association analyses using cladistic-based mapping schemes. Nevertheless, the sensitivity of hierarchical clustering to recombination and gene conversion [Seltman et al., 2003; Templeton et al., 1987; Tzeng et al., 2006] limits single analyses to very small regions. Moreover, not all cladistic mapping methods accept unphased genotype data as input (Table I).

A second approach to replacing the coalescent is a spatial clustering algorithm, wherein haplotypes are grouped according to a distance metric, which simplifies the genetic model even further. Again, the success of these methods depends critically on how closely the clustering technique recreates the natural history of the population. Molitor et al. [2003, 2005] clustered haplotypes deterministically by assignment to one of a set of ancestral central haplotypes, using the length measure about a proposed trait locus as a distance metric. This approach has been refined by Morris [2005, 2006] and by Waldron et al. [2006], who employed measures of similarity that modified the length measure to favor similarities in rare alleles, using weight functions different from that of Durrant et al.

TABLE I. Previously published association mapping methods featuring haplotype clustering

Reference(s)	Program	Unphased data	WG scalable	Quant. data
Bardel et al. [2005; 2006]	ALTree	No	No	No
Browning and Browning [2007a, b]	BEAGLE	No	Yes	Yes
Durrant et al. [2004]	CLADHC	No	Yes	No
Li and Jiang [2005] and Li et al. [2006]	HapMiner	No	Yes	Yes
Liu et al. [2001]	BLADE	Yes	No	No
McPeek and Strahs [1999]	DHSMAP	No	No	No
Molitor et al. [2005]		Yes	No	No
Morris et al. [2002, 2004]	COLDMAP	Yes	No	No
Morris [2005, 2006]	GENEBPM	Yes	No	No
Seltman et al. [2003]	EHAP	Yes	Yes	Yes
Toivonen et al. [2000]	HPM	No	Yes	No
Tzeng [2005] and Tzeng et al. [2006]		Yes	Yes	Yes
Waldron et al. [2006]		No	Yes	No
Yu et al. [2004]		Yes	No	No
Zöllner and Pritchard [2005]	LATAG	No	No	Yes

Program, program name; Unphased data, program accepts genotypes without phase (haplotype) data; WG scalable, scalable for whole-genome analyses; Quant. data, program will analyze continuous trait data.

[2004]. In addition, Yu et al. [2004] grouped haplotypes by constructing nested subsets, based on yet another variation on the length measure. Despite the extreme simplification of the inheritance model, computational demands from the Markov chain Monte Carlo framework make whole-genome association mapping using these techniques infeasible. In a variation on the spatial clustering theme, Browning and Browning [2007a,b] incorporated local LD patterns into haplotype grouping by treating haplotypes as edges in a directed acyclic graph, and merging haplotypes that show similar probabilities for all combinations of alleles that occur downstream within the graph [Browning, 2006].

The distance-based mapping method of Li and Jiang [2005] and Li et al. [2006] is based on data mining techniques. Li and Jiang's [2005] approach uses the core idea that haplotypes carrying trait loci tend to be more similar to each other than haplotypes drawn at random from the population. Haplotypes are clustered by a density-based clustering algorithm [Easter et al., 1996] using a similarity score generalizing two previously described measures of similarity [Tzeng et al., 2003]. Implemented in the program HapMiner, the method is efficient enough for whole-genome studies provided that the sliding window of haplotypes is not too large, and has been adapted for quantitative traits [Li et al., 2006]. HapMiner, like several other cluster-based techniques, is limited in that it requires phased haplotypes as input. Since phase is typically unknown, one possible solution is to infer haplotypes and to use the most likely diplotype for each individual as the "correct" one. However, this solution may cause loss of information when numerous phase resolutions are possible, and the most likely haplotype pair has low probability. Furthermore, in some situations this approach may result in bias [Zhao et al., 2003] and decreased precision [Tanck et al., 2003] in estimated parameters.

In this report, we describe an expansion of the cluster-based association mapping method implemented in HapMiner [Li and Jiang, 2005; Li et al., 2006] to accommodate phase-unknown genotype data. We apply the clustering algorithm from HapMiner to the haplotype score test of Schaid et al. [2002]. We explored the performance and the validity of the new, computationally efficient approach through simulation studies. Our method may be considered a partial generalization of the approach of Tzeng et al. [2006], described above.

METHODS

A CLUSTER-BASED SCORE TEST FOR ASSOCIATION

Our association mapping approach consists of three steps: (1) obtaining posterior probabilities of phased haplotype pairs [S.A.G.E., 2007]; (2) assigning haplotypes consistent with unphased genotypes to clusters [Li and Jiang, 2005]; and (3) assigning all haplotypes within a cluster the same regression parameter in the GLM [Schaid, 2004].

Initially, we calculate probabilities of all possible haplotype pairs for each unphased genotype, using the DECIPHER program in S.A.G.E. [2007]. Probabilities are summed over all individuals in the sample to create haplotype weights for clustering. These weights mimic sample haplotype frequencies, but are actually expected numbers of possible haplotype copies in the sample.

In step (2), we use a modified version of HapMiner to cluster the weighted haplotypes. Instead of assuming that the haplotypes are known for each individual, we use a flexible distance metric [Li and Jiang, 2005] for a set of unique possible haplotypes and their weights. Disease status is not provided to HapMiner, since the score statistic is calculated under the null hypothesis of no haplotype effects. We assign each pair of haplotypes a similarity score $s_{i,j}$, which is a function of the total number of matching alleles (the "counting measure") and the longest interval of continuous matching including the reference marker (the "length measure"). This score generalizes previously described scores [Tzeng et al., 2003]. The similarity score is converted to a distance, $d_{i,j} = (s_{i,i} - s_{i,j})/s_{i,i}$, that is normalized to [0,1]. We then use a modified DBSCAN algorithm [Easter et al., 1996] to group haplotypes. Clusters form in regions of high density. A haplotype is designated a "core" haplotype if enough density, determined by the density threshold MinPts $\in (0,1)$; is located within a given distance ε from it. Haplotypes within this ε neighborhood of a core haplotype are clustered together. "Common" haplotypes with weight greater than a given value $p_{\min} \in (0,1)$ are never clustered with one another. Hence, common, and therefore probably ancient, haplotypes are not improperly clustered. However, haplotypes with weights less than $p_{\min}$ may still be grouped with common haplotypes. We usually set $p_{\min}$ proportional to $1/N$, where N is the number of unique haplotypes detected in the sample. If a haplotype is within distance ε of two common core haplotypes, it is clustered with the one with smaller distance; if the distances are equal, it is grouped with one at random, with probabilities determined by the frequency of the core haplotypes. We chose for this study an exponentially decreasing function for weighting markers contributing to both the similarity and length measures, of the form e^{ax}, where x approximates the physical distance between a given SNP and the reference marker in terms of number of markers (the reference SNP itself is assigned $x = 0$; its immediate neighbors, $x = 1$, etc.), and $a = -0.5$. The reference marker was defined as either the fourth SNP of six, or the fifth SNP of eight, markers in a haplotype.

In some analyses, we allowed $p_{\min}$ to be determined by the Shannon information criterion given the haplotype frequencies in a particular sample [Tzeng et al., 2006]. If k were the number of haplotypes that maximized the Shannon information, $p_{\min}$ was set equal to the frequency of the kth most common haplotype.

In step (3), we perform a score test of association using a GLM approach in the haplotype score test in the HaploStats package [Schaid, 2004; Schaid et al., 2002]. We modified the test to use the cluster assignments to reduce the number of regression parameters. In our approach, posterior probabilities of haplotype pairs are converted to posterior cluster-pair probabilities, and all haplotypes within a cluster are represented by a single model coefficient. The global score test for association is asymptotically distributed, under the null model, as a χ^2 random variable with d.f. equal to the rank of the variance matrix for the score statistic, which may be smaller than the number of clusters. We observed unstable results and

inflation of Type-I error for quantitative traits (data not shown) using the algorithm supplied with HaploStats to calculate the Louis-information-based variance of the score statistic [Louis, 1982]. Consequently, we substituted the calculations from the generalized score test of Boos [1992], using the implementation of Tzeng et al. [2006].

SIMULATING HAPLOTYPE DATA

We generated haplotypes of tightly linked SNPs by a slight modification of the haplotype-extension method of Durrant et al. [2004]. This technique aims to emulate a large population based on a smaller number of founder chromosomes for which phased haplotype data are available. We began with phased haplotypes over a range of 12.3 Mb of chromosome 22 (14.4–26.7 Mb), comprising the 5,000 dimorphic SNPs nearest the p telomere, from all parents of the 30 CEPH trios in the HapMap Project [The International HapMap Consortium, 2005]. The total was 120 presumed unrelated founder chromosomes. First, we randomly chose a genotype at the disease-susceptibility locus for each individual in the simulated sample. For case-control samples, we fixed genotype relative risks (RRs) for genotypes Dd and DD relative to dd, a susceptibility allele frequency p_d, and a disease risk for the dd genotype. Trait genotypes were then chosen using probabilities $\Pr(G \mid \text{case})$ or $\Pr(G \mid \text{control})$, where $G \in \{\text{dd}, \text{Dd}, \text{DD}\}$, as appropriate. For quantitative phenotypes, genotypes were chosen directly from the population under Hardy-Weinberg equilibrium (HWE). Trait values were generated under a normal mixture model. We assumed a genotype mean of 0 for dd individuals and residual variance 1, and chose genotype means for Dd and DD individuals (which are also the genotype effects) according to the desired model.

We then extended haplotypes about the trait alleles as described [Durrant et al., 2004]. In all analyses presented here, we chose a common disease variant with $p_D \sim 0.2$. To increase the diversity in the genetic context, we accepted any SNP within the source data with a minor allele frequency between 0.183 and 0.217 (i.e., between 22 and 26 copies in the set of 120 founder chromosomes) as a potential trait locus. Before analysis, we removed phase information and trait-locus genotypes.

PARAMETER OPTIMIZATION FOR HAPMINER AND POWER ANALYSES

We optimized HapMiner parameter values for $p_{\min}$, ε, and MinPts under various genetic models, and explored the sensitivity of these optima to small changes in the parameters. We typically performed analyses at two different genetic signal strengths. Specifically, we chose model parameters so that the haplotype-only score test would yield roughly 40% (''low'') or 75% (''high'') power at a significance level (α) of 0.05. For studies at low power, we chose either a multiplicative binary-trait model with RR = 1.6 for each susceptibility allele D, or an additive quantitative trait with an allele effect of −0.6 for each D allele. In the high-power model, we either set the allele RR to 1.95 for a multiplicative binary trait, or the allele effect to −0.8 for an additive quantitative trait. Each study simulated 1,000 data sets, each of which comprised either 200 cases and 200 controls for a binary trait, or 200 total individuals for a quantitative trait. Haplotypes contained either six or eight SNPs, not including the trait locus.

Power to detect association was estimated by the proportion of data sets yielding a significant result. Type-I error was measured as the power to detect association from data sets generated under the null model in which the trait locus had no effect on the phenotype. In studies of power vs. haplotype diversity, data for assessing Type-I error were prepared by permuting phenotype values in each sample relative to the genotypes. We compared power and Type-I error among three tests: the haplo.score function with modified variance calculation (HST), our new cluster-based score test (CST), and the test developed by Tzeng et al. [2006] with clustering based on an evolutionary model (TzST). The HST is identical to the CST without clustering. The score test of Tzeng et al. [2006] was applied using the default settings in the R script available from the authors' Web site (http://www4.stat.ncsu.edu/~jytzeng/Softwares/Hap-Clustering/R/).

RESULTS

We explored the properties of the CST, built on HapMiner's clustering algorithm, and compared our test to two related score tests: the modified haplotype-based HST of Schaid et al. [2002] and the cluster-based TzST based on the evolutionary clustering algorithm [Tzeng et al., 2006].

OPTIMIZING HAPMINER PARAMETERS

Maximizing the performance of our new cluster-based score test was primarily a matter of finding the optimal values for the parameters governing the clustering algorithm in HapMiner. We identified a range of values for HapMiner parameters $p_{\min}$, ε, MinPts that provided near-maximum power for both binary- and quantitative-trait data at two different signal strengths. Specifically, power was optimal or near optimal at $p_{\min} = 1/2N$ or $1/3N$, $\varepsilon = 0.4$–0.5, and MinPts = 0.25. The performance of the test, moreover, changed little with small deviations away from the best parameter values.

Power to detect association with a multiplicative binary trait, using haplotypes of six closely spaced markers, was maximized at moderate levels of $p_{\min}$ and at relatively large values of ε (Fig. 1). Restrictions on clustering common haplotypes were clearly required for maximum power. Without such restrictions ($p_{\min} = 1$), performance was relatively poor (data not shown), and in the studies at high power (panels C and D), the CST underperformed the HST at higher ε. This performance likely reflects ''over-clustering,'' in which widely diverged haplotypes are inappropriately clustered, thereby merging disease-susceptible and nonsusceptible lineages. Setting $p_{\min} = 1/N$, where N is the number of distinct haplotypes in the sample, and $\varepsilon = 0.4$ maximized performance for the low-power data (panels A and B) but yielded considerably inferior power at $\alpha = 0.05$ with a strong trait locus (panel C). With $p_{\min} = 1/2N$, $\varepsilon = 0.4$, power was optimal at high power and $\alpha = 0.01$ (panel D), and was very nearly optimal under the other conditions examined (panels A–C). Although maximum performance at high power and $\alpha = 0.05$ was achieved with $p_{\min} = 1/3N$ (panel C), this level

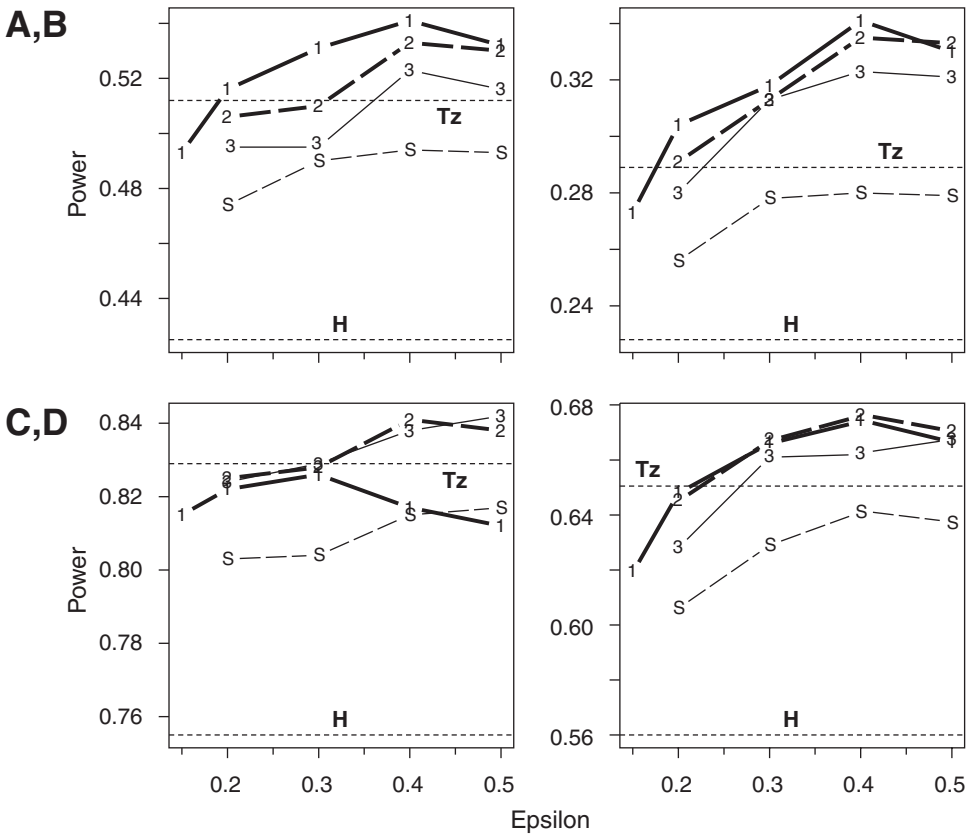

Fig. 1. Optimizing power to detect association with a multiplicative binary-trait locus. Power from the score test was estimated over 1,000 independently generated data sets of 200 cases and 200 controls. In every panel, thin-dotted lines marked as H and Tz indicate power from the HST and the TzST, respectively. Power for the CST is plotted over ranges of ε by different lines and symbols representing values of p_{min}: $p_{min} = 1/N$ (N = number of distinct haplotypes), thick solid line, symbol = 1; $p_{min} = 1/2N$, thick-dashed line, symbol = 2; $p_{min} = 1/3N$, thin solid line, symbol = 3; p_{min} determined by the Shannon information criterion, thin-dashed line, symbol = S. (A, B), Power to detect association under the "low-power" model at nominal Type-I error of 0.05 (A) and 0.01 (B). (C, D), Power to detect association under the "high-power" model at nominal Type-I error of 0.05 (C) and 0.01 (D). The range of the y-axis was determined by the range of performance, and therefore is different in each panel.

of p_{min} returned suboptimal results at low power. Finally, except when $p_{min} = 1$, the Shannon information-guided p_{min} consistently performed very poorly compared to the analyses that more aggressively clustered haplotypes.

At $\varepsilon > 0.3$, both power and extent of clustering were essentially independent of the value of MinPts (data not shown). Consequently, we set MinPts to its default value of 0.25 in our other studies. Our CST slightly outperformed the TzST at optimal parameter values, in an increase of power of 3–5% when analyzing a trait locus with moderate effect (panels A and B) and of 1–3% with a strong trait locus (panels C and D).

The optimal parameter values were similar for an additive quantitative-trait locus (QTL) compared with the results for a binary trait, though not identical (Supplementary Fig. A). The parameter settings $p_{min} = 1/2N$, $\varepsilon = 0.5$ yielded the best overall power: they optimized power at $\alpha = 0.01$ at both high and low power, and tied for greatest power at $\alpha = 0.05$. The TzST performed better, relative to the CST, on quantitative data than on binary data. The best parameter values in HapMiner conferred power less than 2% greater than that of the TzST for a low-power trait locus. The methods were essentially equally powerful for the high-power locus.

The optimal parameter values were similar for eight-marker haplotypes compared with six-marker haplotypes (data not shown), although due to elevated haplotype diversity both cluster-based tests outperformed the HST by a greater margin. On the whole, power to detect association was greater on six-marker haplotypes than on eight-marker haplotypes. This finding most likely reflects reduced correlation among the markers and greater diversity within the longer haplotype.

DIMENSION REDUCTION VS. POWER

Density-based clustering substantially reduced the dimensions of the CST relative to the HST in analyses of a simulated multiplicative binary trait (Fig. 2). As expected, reducing p_{min}, the minimum haplotype frequency at which haplotypes are not clustered, raises the mean d.f. by reducing the extent of grouping. Within a single value of p_{min}, the average number of d.f. fell with increasing ε, reflecting the larger radius of the neighborhoods within which haplotypes can be grouped with a

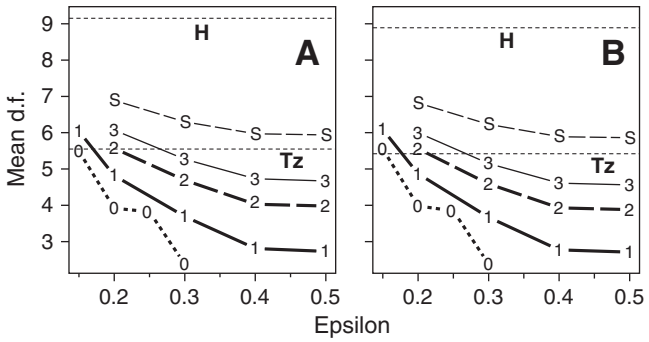

Fig. 2. Mean degrees of freedom (d.f.) of the score test from 1,000 data sets of 200 cases and 200 controls, under the "low-power" (A) and "high-power" (B) scenarios. Lines and symbols for $p_{min} = 1/N$, $1/2N$, $1/3N$ and determined by the Shannon information criterion are as in Figure 1; $p_{min} = 1$ is represented by a thick-dotted line and the symbol 0. Thin-dotted lines indicate mean d.f. of the HST (H) and the TzST (Tz).

"core" haplotype (see the Methods section). At $p_{min} = 1$, indicating no frequency restrictions on clustering, all the haplotypes were grouped into a single cluster with nearly every simulation when ε was set above 0.3; these results are not shown in Figure 2. Otherwise, the d.f. reached a plateau as ε was elevated to 0.5, indicating that every haplotype with frequency less than p_{min} had been grouped with a common haplotype. The median value of N in these analyses was 9 (first and third quartiles = 7 and 13, respectively). Thus, values of p_{min}, whenever p_{min} was defined in terms of $1/N$, were usually between 0.02 and 0.15. For example, the median value of p_{min}, when $p_{min} = 1/2N$, was 0.055 (first and third quartiles = 0.038 and 0.071, respectively). Haplotypes with frequencies below this range could be considered rare enough to warrant clustering with more common haplotypes. Results were nearly identical for data sets generated under low- and high-power trait models (compare Fig. 2, panels A and B). Under optimal HapMiner parameters, our approach clustered haplotypes more extensively than did the evolutionary clustering method of Tzeng et al. [2006], as shown by the reduced average d.f. in the score test. A reduction in the number of predictor variables generally conferred a slight advantage in power under optimal conditions, but clustering to fewer than about 2.5 groups, on average, rapidly diminished power, probably due to improper clustering of benign and deleterious haplotypes (Supplementary Fig. A and data not shown). Power was not strictly a function of the extent of clustering, even near the optimum.

POWER AS A FUNCTION OF HAPLOTYPE DIVERSITY

The analyses heretofore described were performed on data sets with a wide range of haplotype diversity. We next examined the performance of the cluster-based score test on data specifically containing low (5–8 unique haplotypes), medium (9–12), and high (13 and up) haplotype diversity. Based on our previous results, we focused on the three overall best-performing levels of p_{min}: $1/N$, $1/2N$, and $1/3N$.

Power to detect a multiplicative binary-trait locus varied only modestly with choice of p_{min}, except at high power and low haplotype diversity, in which case the setting $p_{min} = 1/N$ performed significantly worse than the other two values (Fig. 3 and Supplementary Fig. B).

Overall, power of both HST and cluster-based tests dropped uniformly with increasing haplotype diversity, suggesting that the genetic architecture in regions of high diversity contains complexity that neither clustering scheme can adequately simplify (Fig. 3 and Supplementary Fig. B, compare low, medium, and high columns in each row). This observation is not surprising, as a causal mutation with susceptibility allele frequency of about 0.2 in a region of high complexity is expected to occur on several different haplotypic backgrounds.

In contrast, the relative *improvement* in performance with clustering compared with no clustering was the greatest at high haplotype diversity, even when overall power was modest. Here, haplotype grouping enhanced performance by 11–13% at low power and by 14–18% at high power. The results of Figure 3 and Supplementary Figure B, furthermore, show a general trend in favor of more aggressive clustering in regions of greater haplotype complexity. Setting p_{min} to $1/N$ invariably yielded the greatest power at high diversity, at both significance levels and at both high and low overall power, resulting in an average number of clusters between 4.2 and 5.4. Results were less consistent at low diversity, although $p_{min} = 1/2N$ proved most versatile in delivering strong results, with mean cluster number from 2.6 to 3.0. In the middle range of diversity, performance varied little with choice of p_{min}.

The CST performed best relative to the TzST at higher haplotype diversity and at modest power. With the best HapMiner parameter settings, the CST conferred greater power than the TzST, with a difference of 4–6%, when analyzing a low-power genetic model in the presence of medium or high diversity. On the other hand, in the presence of a strong trait locus and low diversity, the CST and TzST (with optimal parameters) performed about equally, with a difference in power less than 0.01.

The results from continuous data, with an additive QTL, largely paralleled those from the binary data, with a few exceptions (Supplementary Fig. C). The greatest difference was that improvement in power with both the CST and TzST, relative to the HST, was greater here than for the binary trait (compare Fig. 3 and Supplementary Fig. B with Supplementary Fig. C). As in our initial studies of power at all diversity levels (Figs. 1 and 3 and Supplementary

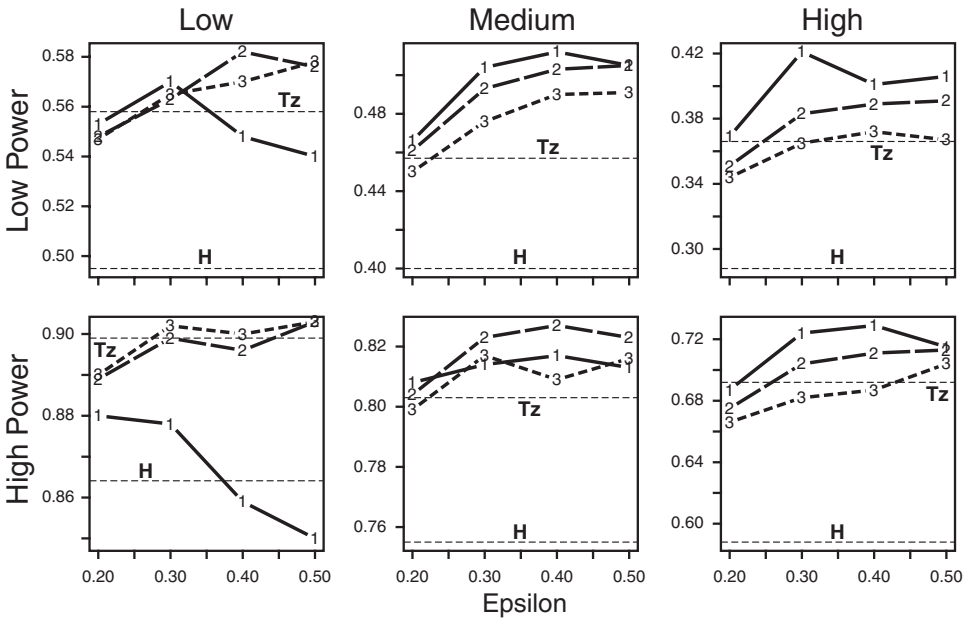

Fig. 3. Power to detect association at the 0.05 significance level with a multiplicative binary-trait locus, stratified by haplotype diversity. For each panel, 1,000 data sets of 200 cases and 200 controls were analyzed. Panels are indicated in terms of haplotype diversity (low, medium, or high) and simulation model (low or high power). In all panels, power of the HST and the TzST is depicted as in Figure 1, and power of the CST is shown as follows: $p_{min} = 1/N$, solid lines, symbol = 1; $p_{min} = 1/2N$, long-dashed lines, symbol = 2; $p_{min} = 1/3N$, short-dashed lines, symbol = 3.

Fig. B), we observed that the CST performed slightly less well relative to the TzST on quantitative data than on binary data (compare Supplementary Fig. C with Fig. 3 and Supplementary Fig. B). Notwithstanding that power at high diversity was consistently several percent greater using the CST with optimal parameters (Supplementary Fig. C, column 3), it showed no improvement over the TzST at low diversity (column 1).

The combined data of Figure 3 and Supplementary Figures B and C show that, although the ideal choice of p_{min} varies directly with haplotype diversity, the precise mathematical relationship is obscure. This overall trend, as well as preliminary studies of the method, discouraged us from fixing p_{min} as a constant, i.e., independent of the number of observed unique haplotypes.

POWER COMPARISONS AT PRESELECTED PARAMETER VALUES

Our approach appears to be limited by the large number of HapMiner parameters over which power must be optimized. However, we have shown above that certain combinations of parameters deliver near-maximum power over a range of conditions tested, in which the strength of the association and the haplotype diversity were varied. To make a ''fair'' comparison among the HST, CST, and TzST, we chose one set of parameter values ($p_{min} = 1/2N$, $\varepsilon = 0.4$) for the CST on binary traits, and another ($p_{min} = 1/2N$, $\varepsilon = 0.5$) for quantitative traits, and assessed power over a range of trait-locus strengths under various genetic models.

Regardless of the mode of inheritance of the trait, the CST consistently outperformed the other two tests at moderate power. Figure 4 presents comparisons of power of the three methods under three types of binary-trait-locus models: multiplicative, dominant, and recessive. Clustering raised performance over the entire range of power under every model (Fig. 4, CST vs. HST). The increase was at a maximum when power was near 50%, as may be expected, but definite improvement was also observed for all disease models when the power of the haplotype method was over 80%. The CST also yielded greater power, by several percent, than the TzST in the range of 40–60% power, but this advantage disappeared at 80% power and above.

With binary-trait data, the CST performed best relative to the HST and the TzST when a recessive trait locus was simulated (Fig. 4, panels E and F), yielding as much as 14% more power than the HST, and 5% more power than the TzST, and performing better than the other two approaches over a range from about 20 to 80% power (panel E). Approximately the same advantage of the CST was observed at $\alpha = 0.01$ (panel F). We noted similar trends in analyses of multiplicative (panels A and B) and dominant (panels C and D) binary-trait loci, although the improvement in power with the CST was slightly smaller in magnitude, with an approximately 10% increase near 50% power.

The CST also fared better at moderate power, relative to the other tests, in a comparison of power to detect QTLs (Supplementary Fig. D), raising the power some 15% over the HST and 2–4% over the TzST. At higher power, the CST and the TzST performed about equally well, and near 80% power both heightened detection of the trait locus by about 6%. In summary, whereas the relative power to detect association by these score tests varies with the type of trait and the mode of trait-locus inheritance, density-based clustering of haplotypes consistently improved performance, and compared well with the evolutionary clustering method of Tzeng et al. [2006].

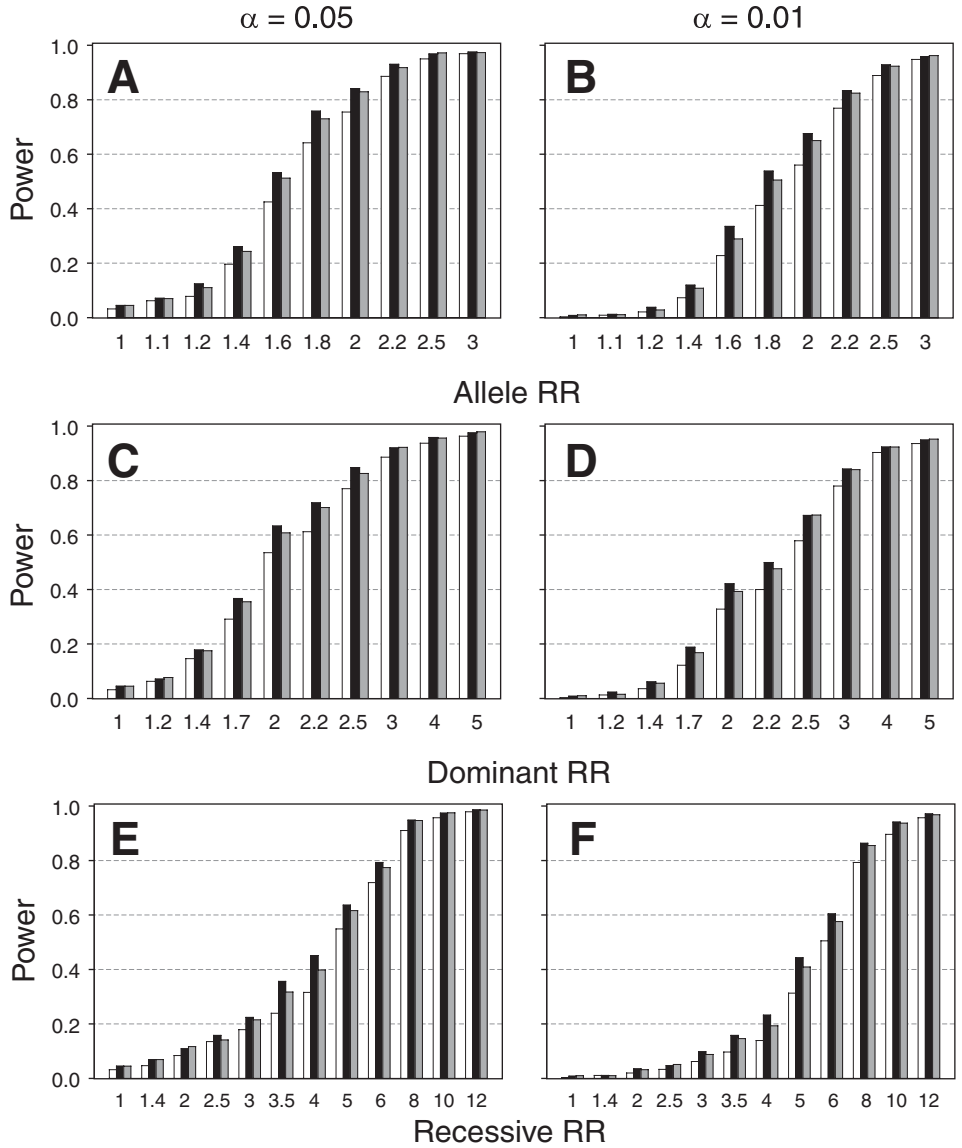

Fig. 4. Comparison of power to detect association with binary-trait data using the HST, the CST, and the TzST. Each triplet of bars represents power estimates from 1,000 data sets of 200 cases and 200 controls. Power is shown at the 0.05 (left column) and 0.01 (right column) significance levels, as a function of signal strength for a multiplicative (A, B), a dominant (C, D), and a recessive (E, F) trait locus. The frequency of the disease allele D is 0.2 throughout. Allele RR, relative risk (RR) associated with each D allele at the trait locus; Dominant RR, RR associated with trait-locus genotype Dd or DD; Recessive RR, RR associated with genotype DD. White bars, HST; black bars, CST; gray bars, TzST.

TYPE-I ERROR

Measurements of Type-I error at near-optimal parameter values show that the CST is valid for use with either binary (Table II) or quantitative (Table III) data. Both the CST and the HST returned false-positive rates within sampling error of the nominal 0.05 and 0.01 Type-I error for a multiplicative binary trait at either low or high power, at all three levels of haplotype diversity described earlier (Table II). Type-I error of the CST was also appropriate when considering quantitative-trait data (Table III). However, the HST was significantly conservative for continuous data, except at high power and low haplotype diversity. The HST grew more conservative at greater haplotype diversity, especially at $\alpha = 0.01$. We could not explain this phenomenon, nor why it was characteristic only of the haplotype-based test, although we did observe the same effect in a binary trait adjusted for covariates [Igo et al., 2007].

DISCUSSION

Our results show that a novel test using density-based clustering substantially enhances power to detect association under an array of genetic models, whether evidence

TABLE II. Type-I error of haplotype- and cluster-based tests as a function of haplotype diversity, with binary trait

		Low Power		High Power	
Diversity	Analysis	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Low	Haplo.	0.038	0.012	0.042	0.014
	Clusters	0.059	0.012	0.038	0.012
Medium	Haplo.	0.039	0.009	0.036	0.006
	Clusters	0.051	0.011	0.036	0.007
High	Haplo.	0.045	0.009	0.040	0.008
	Clusters	0.046	0.006	0.050	0.012

Haplo., HST; Clusters, CST; α, nominal Type-I error. See the Methods section for definitions of low, medium, and high diversity and of low and high power.

TABLE III. Type-I error of haplotype- and cluster-based tests as a function of haplotype diversity, with quantitative trait

		Low Power		High Power	
Diversity	Analysis	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
Low	Haplo.	*0.031*	*0.005*	0.039	0.006
	Clusters	0.041	0.009	0.058	0.011
Medium	Haplo.	*0.033*	*0.000*	*0.023*	*0.000*
	Clusters	0.061	0.008	0.039	0.007
High	Haplo.	*0.022*	*0.001*	*0.018*	*0.001*
	Clusters	0.044	0.009	0.046	0.006

Entries in italics are significantly different from the nominal level, as determined by an exact test using a binomial probability distribution. Designations are as in Table II.

for association is modest or strong, with valid Type-I error rates. The clustering algorithm groups haplotypes likely to share common recent ancestry. This technique reduces the complexity of the score test for association developed by Schaid et al. [2002]. Improvement in power to detect QTLs was somewhat greater than that to detect binary causative loci, though the reason for this was not clear. While a major strength of our cluster-based approach is the flexibility in the HapMiner clustering algorithm, we identified sets of HapMiner parameters that delivered near-optimum performance under several inheritance patterns and over a broad range of haplotype diversity. The GLM framework enables analysis of several types of data—binary, continuous, and Poisson (count) data foremost among them—and also, when needed, incorporates environmental covariates. In addition, the method is computationally efficient: analyses of 1,000 data sets of 200 cases and controls and 6 SNPs required only about 35 min on a 2.6-GHz AMD processor running Linux. For short haplotypes (less than 10 markers), the speed of the new method is comparable to that of the score test of Tzeng et al. [2006], whereas analysis of longer haplotypes tends to slow HapMiner (data not shown). However, a sliding window of six to eight SNPs is expected to cover most regions of strong LD in a genome-wide scan comprising approximately 500,000 markers [The International HapMap Consortium, 2007]. Our approach can accommodate missing data, at the expense of additional computational complexity. The effect of missing data on the computational burden, and on the performance of the clustering algorithm, is a topic for future investigation.

We expect our approach to work well in the presence of both allelic and locus heterogeneity. The simulation method here models allelic heterogeneity by placing disease alleles on a variety of haplotype backgrounds. We did not explicitly simulate locus heterogeneity, but the binary-trait model allowed sporadic cases, and most quantitative-trait models showed considerable overlap in the trait values of the three genotypes at the trait locus.

Our studies revealed that although optimizing HapMiner parameters is potentially laborious, a certain range of parameter values seems to work well in several contexts. Such ''all-purpose'' parameter settings are necessary for the success of genome-wide association mapping using this method. In certain cases, clustering under different values of $p_{\min}$ yielded very similar power despite considerably different mean d.f. The overall haplotype diversity affected optimal parameter choice more strongly, but in a predictable way.

We do not attempt to localize causative variants. The cluster-based mapping approaches that do, however, are too computationally intensive to be useful for scans of large genomic regions [e.g., Bardel et al., 2005, 2006; Zöllner and Pritchard, 2005]. However, individual clusters with highly significant effect under the score test may be good candidates for harboring causal variants. Our simulations, in which the causative locus was omitted from the haplotype, provide a realistic simulation of a genome-wide association analysis using a random (agnostic) or tag SNP set. We did not conduct simulations in which genotypes of a single causative SNP were available because, in this situation, single-SNP association is expected to be more powerful.

Because our simulation studies required analyses on thousands of unique data sets, we could not use a weighting function for haplotype similarity that incorporated potentially useful information about physical distance or LD between the reference SNP and other SNPs in the haplotype. HapMiner, however, does allow weighting functions based on either map distance or LD. We chose the exponential function, assigning greater weight to more central markers. This weighting scheme has the advantage of being robust to recombination. In practice, the ideal number of SNPs in the haplotype and the best weight function may depend on the density of the sampled SNPs and the LD structure of the region of interest. In a genome-wide scan, these properties will vary widely, and therefore a generic, simple weighting function like the exponential is likely to work best. Having focused on features of the CST relevant to genome-wide association mapping, we did not evaluate, in this study, its performance in a fine-mapping context in which the LD structure would play an important role. We expect, nonetheless, that the CST would perform favorably, relative to HapMiner, when LD is used as the criterion for clustering, since a major advantage of the CST is that, unlike HapMiner, it does not require phased haplotype data.

As with any statistical approach, our clustering method is limited in several respects. We have sacrificed some

accuracy in modeling to reduce the computational burden. Thus, the approach may not be ideal with data from populations markedly different from the Utah CEPH families, who were the source of the haplotype data from which genetic data were simulated. There is no clear-cut strategy for determining the proper number of SNPs to include in a haplotype. Several criteria are commonly used, such as map density and intermarker correlation. Posterior haplotype probabilities are calculated under the assumption of HWE. The approach is not expected to offer great power to detect rare variants, in part because haplotypes are parameters. Therefore, our approach creates an additive haplotype effect on the transformed trait (where the transformation depends on the relevant GLM). Indeed, we were unable to detect a rare disease susceptibility locus simulated for Genetics Analysis Workshop 15 (GAW15) [Igo et al., 2007]. Posterior haplotype probabilities are calculated under the assumption of HWE. We did not test here the sensitivity of our approach to departures from HWE. However, like the haplotype-based score test in HaploStats [Schaid, 2004], the CST is robust to the severe departure from HWE observed in the vicinity of the HLA-DRB1 locus in the simulated data from GAW15, in which there was an excess of homozygosity [Igo et al., 2007].

In the technique described here, we cluster and predict haplotypes independently of trait status. Our method improved performance most dramatically in the presence of extreme haplotype diversity, with slight improvement at low diversity. Our method also offers, for binary traits, a slight advantage in power over the Tzeng et al. approach. To estimate cluster effects, we would ideally account for the trait status, as in the haplo.glm regression-based association test in HaploStats [Lake et al., 2003; Schaid, 2004]. Using the framework of haplo.glm would also enable us to model gene-environment interaction. Because clustering in this context will relieve some of the problems of large numbers of coefficients necessary to model all haplotype-covariate interactions, we expect that our clustering algorithm will enhance the power of this more sophisticated test, as well. This refinement of the CST is a subject for further research.

ACKNOWLEDGMENTS

We thank Dan Baechle for his programming expertise. R.P.I. was supported by NIH grant HL07567. J.L. was supported in part by NIH/NLM grant 008911 and a startup fund from Case Western Reserve University. K.A.B.G. was supported by a grant from the National Center for Research Resources (RR03655), and by a grant from the NIH/NIDDK (P30 DK027651). Some of the results in this paper were obtained using the program package S.A.G.E., which is supported by a US Public Health Service Resource Grant (RR03655) from the National Center for Research Resources.

REFERENCES

Akey J, Jin L, Xiong M. 2001. Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291–300.

Bardel C, Danjean V, Hugot J-P, Darlu P, Génin E. 2005. On the use of haplotype phylogeny to detect disease susceptibility loci. BMC Genet 6:24.

Bardel C, Danjean V, Génin E. 2006. ALTree: association detection and localization of susceptibility sites using haplotype phylogenetic trees. Bioinformatics 22:1402–1403.

Boos DD. 1992. On generalized score tests. Am Stat 46:327–333.

Bourgain C, Génin E, Quesneville H, Clerget-Darpoux F. 2000. Search for multifactorial disease susceptibility genes in founder populations. Ann Hum Genet 64:255–265.

Browning SR. 2006. Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78:903–913.

Browning BL, Browning SR. 2007a. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol 31:365–375.

Browning SR, Browning BL. 2007b. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097.

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB. 2001. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361–364.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75: 35–43.

Easter M, Kriegel HP, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, OR: AAAI Press. p 226–231.

Fallin D, Schork NJ. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959.

Igo Jr RP, Londono D, Miller K, Parrado AR, Quade SRE, Sinha M, Kim S, Won S, Li J, Goddard KAB. 2007. Density-based clustering in haplotype analysis for association mapping. BMC Proc 1:S27.

The International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.

Kingman JFC. 1982. The coalescent. Stochastic Process Appl 13:235–248.

Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered 55: 56–65.

Li J, Jiang T. 2005. Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics 21:4384–4393.

Li J, Zhou Y, Elston RC. 2006. Haplotype-based quantitative trait mapping using a clustering algorithm. BMC Bioinformatics 7:258.

Liu JS, Sabatti C, Teng J, Keats BJB, Risch N. 2001. Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res 11:1716–1724.

Louis TA. 1982. Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 44:226–233.

McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875.

Molitor J, Marjoram P, Thomas D. 2003. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384.

Molitor J, Zhao K, Marjoram P. 2005. Fine mapping—19th century style. BMC Genet 6:S63.

Morris AP. 2005. Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modeling of haplotypes. Genet Epidemiol 29:91–107.

Morris AP. 2006. A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. Am J Hum Genet 79:679–694.

Morris AP, Whittaker JC, Balding DJ. 2002. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 73:1368–1384.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953.

S.A.G.E. 2007. Statistical Analysis for Genetic Epidemiology, version 5.4. http://darwin.cwru.edu/sage/.

Schaid DJ. 2004. Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348–364.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434.

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

Shannon CE. 1948. A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656.

Tanck MW, Klerkx AH, Jukema JW, De Knijff P, Kastelein JJ, Zwinderman AH. 2003. Estimation of multilocus haplotype effects using weighted penalised log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. Ann Hum Genet 67:175–184.

Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehyrogenase activity in Drosophila. Genetics 117:343–351.

Thomas DC, Stram DO, Conti D, Molitor J, Marjoram P. 2003. Bayesian spatial modeling of haplotype associations. Hum Hered 56:32–40.

Toivonen HTT, Onkamo P, Vasko K, Ollikainen V, Sevon P, Mannila H, Herr M, Kere J. 2000. Data mining applied to linkage disequilibrium mapping. Am J Hum Genet 67:133–145.

Tzeng J-Y. 2005. Evolutionary-based grouping of haplotypes in association analysis. Genet Epidemiol 28:220–231.

Tzeng J-Y, Devlin B, Wasserman L, Roeder K. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891–902.

Tzeng J-Y, Wang C-H, Kao J-T, Hsiao CK. 2006. Regression-based association analysis with clustered haplotypes through use of genotypes. Am J Hum Genet 78:231–242.

Waldron ERB, Whittaker JC, Balding DJ. 2006. Fine mapping of disease genes via haplotype mapping. Genet Epidemiol 30:170–179.

Yu K, Martin RB, Whittemore AS. 2004. Classifying disease chromosomes arising from multiple founders, with application to fine-scale haplotype mapping. Genet Epidemiol 27:173–181.

Zhao H, Pfeiffer R, Gail MH. 2003. Haplotype analysis in population genetics and association studies. Pharmacogenomics 4:171–178.

Zöllner S, Pritchard JK. 2005. Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169: 1071–1092.