ARTICLE

Multipoint Linkage-Disequilibrium Mapping with Haplotype-Block Structure

Maoxia Zheng and Mary Sara McPeek

The HapMap Project is providing a great deal of new information on high-resolution haplotype structure in various human populations. This information has the potential to greatly increase the power of association mapping for a fixed amount of genotyping. A number of methods have been proposed for the identification of haplotype blocks, common haplotypes, and tagging single-nucleotide polymorphisms. Here, we build on this work by developing novel methods for case-control multipoint linkage-disequilibrium (LD) mapping that gain power and speed by making explicit use of the inferred block structure. Specifically, we developed a virtual-variant approach that uses the haplotype-block information to greatly increase power for detection of untyped common variants associated with a trait. Because full multipoint LD mapping can be slow, we exploited the haplotype-block information to develop a fast single-block multipoint mapping method. Our methods are appropriate for genotype data and take into account the uncertainty in phase. We describe the methods in the context of case-parents trios, although they are also applicable to unrelated cases and controls. Our simulations indicate that the most important gains from taking into account the haplotype-block structure at the analysis stage of multipoint LD mapping come from (1) greatly increased power to detect association with untyped variants and (2) greatly improved localization of untyped variants associated with the trait. More-modest gains are obtained in improving power to detect association with a variant that is typed with a moderate amount of missing data. The methods are applied to a Crohn disease data set.

On the basis of the suggestion of haplotype-block structure in various human populations,[1–4] the International HapMap Project,[5] aimed at determining the common patterns of DNA sequence variation in the human genome for association studies, has provided a large amount of new information about high-resolution haplotype structure. Although haplotype-block structure might not hold generally across the whole genome,[6] it has the potential to be useful for association mapping[7,8] in the regions that show substantial haplotype-block structure. There is much previous work on identification of blocks, common haplotypes, and haplotype tagging SNPs (htSNPs).[1,2,8–16] The purpose of the present study was to build on that previous work by developing mapping methods specifically tailored to take advantage of the information about haplotype-block structure.

For the purposes of our methods, a haplotype block could consist of a set of SNPs across which there is only a relatively small number of common haplotypes. We do not require any conditions on pairwise linkage disequilibrium (LD) between SNPs in a block. We impose no strict upper limit on the number of common haplotypes in a block, but the power and speed of the method are linked to this number being small. In principle, a haplotype block would not necessarily have to be an interval but could be a union of intervals, and haplotype blocks could be overlapping. We assume haplotype blocks and common haplotypes have been identified, and we consider robustness of our methods to the choice of these by comparing results obtained with the use of each of six different haplotype-block algorithms.

We introduce four types of multipoint analysis. Some of the potential advantages of multipoint methods include (1) their use of more of the information in the data when a susceptibility variant in the region is untyped or partially typed and (2) the fact that likelihoods at nearby variants are based on the same data, so they are formally comparable for the purposes of localization. As a result, multipoint methods have the potential to vastly improve localization over single-point methods. Because full multipoint-likelihood methods require a model for background LD and come with a high computational cost, we propose two single-block multipoint methods that use only the marker information in the same haplotype block as the variant being tested. The resulting methods are essentially nonparametric with respect to background LD and have greatly reduced computation time compared with full multipoint analysis, at the cost of some loss of information. In contrast, our two full multipoint-likelihood methods use marker information across multiple blocks in the region of the variant being tested.

Within each of the single-block and full multipoint approaches, we first propose a method that considers association only with typed variants; we call these the "single-block observed variant" (SBOV) and "full multipoint observed variant" (MOV) association methods. However, in practice, because of cost considerations, there will typically be many more untyped than typed variants in a re-

From the Departments of Statistics (M.Z.; M.S.M.) and Human Genetics (M.S.M.), University of Chicago, Chicago
Received July 19, 2006; accepted for publication November 7, 2006; electronically published November 30, 2006.
Address for correspondence and reprints: Dr. Mary Sara McPeek, Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago, IL 60637. E-mail: mcpeek@galton.uchicago.edu
Am. J. Hum. Genet. 2007;80:112–125. 0002-9297/2007/8001-0012$15.00

gion. A major rationale for the HapMap Project and for the selection of tagging SNPs (tSNPs) to identify common haplotypes in a block is the so-called common variant–common disease hypothesis, which suggests that common variants are particularly plausible as susceptibility variants. To detect the presence of an untyped common SNP associated with the trait in a particular haplotype block, we characterize the untyped SNP by partitioning the set of haplotypes within the block into two disjoint subsets. That is, we assume that the SNP allele does not vary within any common haplotype in the block and so, for instance, the SNP might be assumed to have one allele on a particular subset of common haplotypes and another allele on the complement of that subset. We call these pseudo-SNPs "virtual variants" (VVs). Using this characterization, we propose both single-block VV (SBVV) and full multipoint VV (MVV) association methods. We expect that VV methods will extract more information on association with untyped variants but at the cost of a higher penalty for multiple comparisons.

The methods we propose are appropriate for a study design consisting of unrelated cases and controls. We are able to apply our case-control methods to trio data by treating the transmitted haplotypes as samples from cases and the nontransmitted haplotypes as samples from controls. Trio data contain more information about phase than do data from unrelated cases and controls, but, otherwise, we do not make use of the trio structure in our methods.

Simulation studies are performed to compare single-point association analysis, SBOV, SBVV, MOV, and MVV, in terms of power to detect association, accuracy of localization of a causal variant, and robustness to choice of haplotype-block boundaries. The methods are applied to a Crohn disease (CD [MIM 266600]) data set.[2]

Methods

Single-Point Association Analysis

Various types of single-point analyses have been commonly used to test for association of a trait with a marker. Although Pearson's χ^2 test and the transmission/disequilibrium test (TDT)[17] are perhaps more widely used, we consider the likelihood-ratio test, because it allows, in a straightforward way, for incomplete information and is the most easily extendable to multipoint-association analysis. With complete data, the likelihood-ratio and Pearson's χ^2 tests are asymptotically equivalent.

Suppose we have genotype data, on multiple biallelic SNPs, for trios consisting of two parents and an affected offspring. For each SNP, for simplicity, we denote the set of possible alleles by {1,0}. When genotype data are available for a SNP on all three members of a trio, then one can determine unambiguously the numbers of type 1 alleles among the transmitted and among the nontransmitted alleles. The model we use for the likelihood-ratio test assumes that all parental alleles are independent, with transmitted alleles being independent, identically distributed (IID) draws from a Bernoulli (p) distribution and nontransmitted alleles being IID draws from a Bernoulli (q) distribution. The single-point association test is the likelihood-ratio test of the null hypothesis $p = q$ for a given SNP (temporarily designated "the SNP of interest"), versus alternative $p \neq q$, which is based on the genotype data of only the SNP of interest.

Suppose we have complete (unphased) genotype data for n trios. Let n_t and n_u denote the number of type 1 alleles among the transmitted and nontransmitted alleles, respectively. The single-point likelihood-ratio test statistic for the SNP of interest is then

$$T_{\text{single}} = 2\log\left[\frac{\hat{p}^{n_t}(1-\hat{p})^{2n-n_t}\hat{q}^{n_u}(1-\hat{q})^{2n-n_u}}{\hat{q}_0^{n_t+n_u}(1-\hat{q}_0)^{4n-n_t-n_u}}\right], \qquad (1)$$

where $\hat{p} = n_t/2n$ and $\hat{q} = n_u/2n$ are the maximum-likelihood estimators (MLEs) of p and q, respectively, under the alternative hypothesis ($p \neq q$), and $\hat{q}_0 = (n_t + n_u)/4n$ is the MLE of $p = q$ under the null hypothesis. T_{single} is known from classic theory to be asymptotically χ^2_1 under the null hypothesis. Under the assumption that the occurrence of missing genotypes in the data is independent of the true genotypes, one can compute the likelihood-ratio statistic T_{single} even when some genotypes are missing. In that case, we compute T_{single} by using the expectation-maximization (EM) algorithm[18] to maximize the likelihood, instead of by using equation (1).

In some instances, single-point association analysis has been quite successful.[19] However, when a susceptibility variant is untyped or typed only in some individuals, multipoint-association analysis, in which genotypes at nearby markers are taken into account, would be expected to extract more of the information from the data.

Single-Block Association Analysis

Assume haplotype block boundaries and common haplotypes in each block have been defined. We propose SBOV and SBVV, which use the multipoint information within a single haplotype block to detect whether there are typed or, in the case of SBVV, untyped variants within this block that are strongly associated with the trait.

SBOV test for association with typed variants.—The model used for single-block association analysis can be thought of as an extension of the model used for single-point association analysis. As before, for any given SNP, we assume that all parental alleles are independent, with transmitted alleles being IID draws from a Bernoulli (p) distribution and nontransmitted alleles being IID draws from a Bernoulli (q) distribution; as before, we perform a likelihood-ratio test of the null hypothesis $p = q$, versus the alternative $p \neq q$. However, in this case, the likelihood is based on data from all typed SNPs in the same haplotype block with the SNP of interest, whereas, for the single-point test, it was based on only the SNP of interest. We call the resulting likelihood-ratio test statistic "T_{SBOV}."

We describe the model in terms of the unobserved complete data for each trio, which consist of the transmitted and nontransmitted alleles for each parent for each SNP in the same haplotype block with the SNP of interest. We then use the EM algorithm[18] to maximize the likelihood for the case of incomplete data, where the likelihood is summed over all possible haplotypes compatible with the observed genotype data for the trios. Let $k = 1, \ldots, K$ index the markers in the block containing the SNP of interest. Then, for each parent, the transmitted haplotype is a vector of length K, with the kth component equal to the parent's transmitted allele at marker k, and likewise for the nontransmitted haplotype. Suppose d is the index of the SNP of interest. Then

the transmitted and nontransmitted alleles at SNP d are assumed to be IID Bernoulli (p) and IID Bernoulli (q), respectively. To model the background LD, let $P_T(H = h|H_d = h_d)$ denote the probability that a transmitted haplotype H is of type h, given that it matches h at SNP d, where $H,h \in \{1,0\}^K$. We define P_N similarly for nontransmitted haplotypes. Both our null and alternative models assume that $P_T(H = h|H_d = h_d) = P_N(H = h|H_d = h_d)$ for all choices of (h,d), in which case we write this probability as simply $\alpha_{h,d} = P(H = h|H_d = h_d)$. This would hold if, aside from the SNP of interest, there were no susceptibility variant in LD with any marker in the haplotype block; in other words, in the case of a single susceptibility variant in the region. (This still leaves open the possibility of many other susceptibility variants in the genome.)

In principle, for any given choice of d, the model for $\alpha_{h,d} = P(H = h|H_d = h_d)$ is the fully parameterized one, with $2^K - 2$ freely varying parameters (corresponding to the 2^K choices of h subject to two constraints: $\sum_{h:h_d=1} \alpha_{h,d} = 1$ and $\sum_{h:h_d=0} \alpha_{h,d} = 1$). In practice, because of shared ancestry in the evolution of haplotypes, only a small fraction of the 2^K possible haplotype values would be expected to arise in the history of the population when K is large. In fact, the way in which haplotype blocks are constructed ensures that only a small number of haplotypes are present. If complete data were available, this fully parameterized model would therefore present little difficulty, even when K is large, because the number of nonzero background-LD parameter estimates would typically be quite moderate. However, when only genotype data are observed, there can be many possible sets of haplotypes compatible with a given family's genotype data, so many or even all of the background LD parameters would have nonzero maximum-likelihood estimates, but with the majority of these being negligible. When K is large, we therefore perform a preprocessing step in which a subset $S \subset \{1,0\}^K$ of haplotypes is chosen, and the model is restricted so that all haplotypes are assumed to lie in S. For a reasonably large data set, S can be chosen by use of existing software (e.g., PHASE[20]) to infer haplotypes for all individuals in the data and then by use of the set of distinct inferred haplotypes as S. Note that the inferred haplotypes are used only to choose S; once S is chosen, the original genotype data are analyzed, with the phase uncertainty taken into account in the analysis.

We now describe two properties of our model for single-block association analysis. First, the maximized log-likelihoods under the null hypothesis are identical for different choices of the SNP of interest within the same haplotype block. This is a consequence of the fact that, under the null hypothesis, our model for the data is that transmitted and nontransmitted haplotypes for the block are IID draws from a fully parameterized discrete distribution on S, and the same data are used in the likelihoods for different choices of the SNP of interest within the same haplotype block. Furthermore, for each choice of the SNP of interest, the alternative model has one additional free parameter beyond the parameters in the null model, and the same data are used in the alternative likelihoods for different choices of the SNP of interest within the same haplotype block. As a consequence, test results from different SNPs within a block are formally comparable, and so the method is expected to be particularly useful for localization within a block.

The second property is that when genotype data are available for the SNP of interest for all individuals, the SBOV likelihood-ratio test statistic, T_{SBOV}, is equal to T_{single}, the single-point likelihood-ratio test statistic (see appendix A for the outline of the derivation). That is, when there are complete data for the SNP of interest, then the multipoint approach provides no additional information for testing the association of the trait with that SNP. This is clearly a desirable property; in fact, if it did not hold, that would signal a problem with the model. When there are incomplete genotype data for the SNP of interest, the two statistics are, in general, different, with T_{SBOV} expected to provide a more powerful test.

SBVV test for association with untyped variants.—One important aspect of haplotype-block structure is that, in each block, only a relatively small set of tSNPs needs to be typed to distinguish the common haplotypes from each other. Because of the generally high level of background LD within a haplotype block—because of the tree structure for the ancestry of the individuals, with mutations occurring over the history of the population—tSNP genotype data can be expected to provide information on some of the common untyped SNPs in the block. To detect the presence of an untyped common SNP associated with the trait, in a particular haplotype block, we characterize the untyped SNP by partitioning the set of haplotypes within the block into two disjoint subsets. That is, we assume that the SNP allele does not vary within any common haplotype in the block, so, for instance, the SNP might be assumed to have one allele on a particular subset of common haplotypes and another allele on the complement of that subset. We call these pseudo-SNPs "virtual variants."

For example, suppose that when the tSNPs are typed for a particular haplotype block, the set S of haplotypes consists of four distinct "common" (frequency $\geq$0.05) haplotypes, labeled 1, 2, 3, and 4, as well as a collection of "rare" (frequency of each <0.05) haplotypes, labeled r. We consider the VVs corresponding to all two-component partitions of S under the restriction that the rare haplotypes are kept together and are always combined with at least one common haplotype: (1, 234r), (2, 134r), (3, 124r), (4, 123r), (12, 34r), (13, 24r), (14, 23r), (1r, 234), (23, 14r), (24, 13r), (2r, 134), (34, 12r), (3r, 124), or (4r, 123). Here, for instance, (23, 14r) represents a VV that is assumed to have, say, allele 1 on every haplotype labeled 2 or 3 and allele 0 on every haplotype labeled 1, 4, or r.

The SBVV is otherwise identical to the SBOV method. In the notation of the "SBOV test for association with typed variants" section, the SNP of interest, d, is the VV, and the other SNPs in the same haplotype block with it are simply the genotyped tSNPs for that block. We call the resulting likelihood-ratio test statistic "T_{SBVV}." If a VV is found to be strongly associated with the trait and if it does not correspond (modulo rare haplotypes) to a typed SNP, then this could be suggestive of the possibility of untyped variation strongly associated with the trait. In that case, additional resources could be expended to explore more extensively the SNPs in the corresponding block.

One feature that distinguishes SBVV from SBOV is that SBVV typically involves performing more hypothesis tests per block than does SBOV. On the one hand, SBVV has the potential to make better use of the data than does SBOV in the case where there is an untyped susceptibility variant, whereas, on the other hand, there will be a higher price to pay for multiple comparisons. This trade-off is explored in the simulation studies.

Full Multipoint-Association Analysis

Whereas the SBOV and SBVV multipoint methods use only the information on typed SNPs within the same haplotype block as

Table 1. Power at Level $\alpha = .05$ for Site-Specific Tests at a Typed Causal SNP

p[a] and Proportion of Missing p_m	Power of Method				
	Single Point	SBOV		MOV	
		Dense[b]	Tag[c]	Dense	Tag
.42:					
.00	.95	.95	.95	.95	.95
.15	.91	.95	.92	.95	.94
.30	.85	.95	.90	.95	.93
.40:					
.00	.88	.88	.88	.88	.88
.15	.82	.88	.85	.88	.87
.30	.74	.88	.81	.88	.86
.35:					
.00	.49	.49	.49	.49	.49
.15	.42	.48	.45	.49	.48
.30	.37	.48	.41	.49	.47

NOTE—The causal SNP is marker 31 from the CD data set[2] (5th tSNP) with nontransmitted allele frequency $q = 0.27$; 10,000 data sets were simulated.

[a] Allele frequency among transmitted alleles.

[b] Genotype data are simulated on all 103 SNPs of the CD data set.[2]

[c] Genotype data are simulated on 27 tSNPs for the CD data set.[2]

the SNP of interest, the corresponding full multipoint methods use information on typed SNPs across multiple haplotype blocks in the region of the SNP of interest. The full multipoint-association analyses for observed variant (i.e., MOV) and for VV (i.e., MVV) are analogous to SBOV and SBVV, respectively.

We first describe the model for the situation in which we have complete data for each trio. In this context, we consider complete data to consist of the transmitted and nontransmitted haplotypes for each parent for each haplotype block in a specified region around the SNP of interest. Let $t = 1, \ldots, B$ index the haplotype blocks, and let K_t be the number of typed SNPs in block t. Then, for each parent, the transmitted haplotype for the tth block is a vector of length K_t, with the kth component equal to that parent's transmitted allele at the kth marker in the block, and likewise for the nontransmitted haplotype. Let d be the index of the SNP of interest, and let δ be the index of its haplotype block. As in the "SBVV test for association with untyped variants" section, we assume that $P_T(H = h|H_d = h_d) = P_N(H = h|H_d = h_d)$ for all $H,h \in \{1,0\}^K$, where $K = \sum_t K_t$ and H represents a combined haplotype across all blocks. Whereas, for the single-block method, we used a fully parameterized model for $P(H = h|H_d = h_d)$, in the full multipoint method we use a Markov model indexed by block. Thus, we have background LD parameters $\alpha_d, \theta_{2\to1}, \ldots, \theta_{d\to d-1}, \theta_{d\to d+1}, \ldots, \theta_{B-1\to B}$. Here, $\alpha_d = (\alpha_{h_\delta,d}, h_\delta \in S_\delta)$, where S_δ is the sample space for haplotypes in block δ, the block containing the SNP of interest, and $\alpha_{h_\delta,d} = P(H_\delta = h_\delta|H_{\delta,d} = h_{\delta,d})$ represents the conditional probability that a randomly drawn haplotype for block δ is of type h_δ, given that it matches h_δ at SNP d. Thus, α_d has the same meaning as in the "SBVV test for association with untyped variants" section. The parameters $\theta_{t_1\to t_2} = [\theta_{t_1\to t_2}(h_{t_1},h_{t_2}), h_{t_1} \in S_{t_1}, h_{t_2} \in S_{t_2}]$ are the one-step Markov transition probabilities between adjacent haplotype blocks, where $\theta_{t_1\to t_2}(h_{t_1},h_{t_2})$ represents the probability that a randomly drawn haplotype for the entire region has haplotype h_{t_2} in block t_2, given that it has haplotype h_{t_1} in block t_1. Here, S_t represents the sample space for the haplotypes from block t. For an examination of this choice of model, see the "Assessment of Goodness of Fit of the HMM for Background LD" section.

For the incomplete-data case, we use a hidden Markov model (HMM), which allows both likelihood calculation and maximization of the likelihood over the parameters. For each trio, we consider the hidden Markov chain

$$\{X_t, t = 1, \ldots, B\} = \{(X_t^{MT}, X_t^{MN}, X_t^{FT}, X_t^{FN}), t = 1, \ldots, B\},$$

where X_t^{MT} denotes the mother's transmitted haplotype in block t, X_t^{MN} denotes the mother's nontransmitted haplotype in block t, X_t^{FT} denotes the father's transmitted haplotype in block t, and X_t^{FN} denotes the father's nontransmitted haplotype in block t. Here, $\{X_t^{MT}\}$, $\{X_t^{MN}\}$, $\{X_t^{FT}\}$, and $\{X_t^{FN}\}$ are four independent Markov chains, parameterized as described above, with the parameters for the transmitted and nontransmitted haplotypes differing only in the frequency of the SNP of interest (p for the transmitted haplotypes and q for the nontransmitted haplotypes). The observed process is taken to be $\{Y_t\}$, where Y_t denotes the (unphased) genotype data in block t for this family, with missing data permitted. Because of the Markov structure, the HMM can be efficiently fit to data via the EM procedure,[18] by an extension of the Baum[21] algorithm for HMMs.

The two properties described in the "SBOV test for association with typed variants" section generalize to the full multipoint-likelihood method—namely, (1) hypothesis-testing results are formally comparable for different choices of the SNP of interest across the region (not just within the same haplotype block), so they are appropriate to use for localization of a causal SNP, and (2) when genotype data are available for the SNP of interest for all individuals, the full multipoint likelihood-ratio test statistics for MOV and MVV—T_{MOV} and T_{MVV}—are both equal to T_{single}, the single-point likelihood-ratio test statistic. The multipoint statistics provide additional information when there are missing data at the SNP of interest or when the SNP is untyped.

Table 2. Power of Regionwide Tests with Use of tSNPs When Causal SNP Is Untyped

Causal SNP Model and p	Power of Method (SE)				
	Single Point	SBOV	MOV	SBVV	MVV
1[a]:					
.35	.84 (.03)	.83 (.03)	.81 (.03)	.91 (.02)	.99 (.01)
.30	.58 (.03)	.58 (.03)	.61 (.03)	.82 (.03)	.88 (.02)
.25	.30 (.03)	.30 (.03)	.30 (.03)	.37 (.03)	.43 (.04)
2[b]:					
.50	.64 (.03)	.65 (.03)	.64 (.03)	.93 (.02)	.98 (.01)
.45	.43 (.04)	.44 (.04)	.44 (.04)	.76 (.03)	.90 (.02)
.40	.26 (.03)	.26 (.03)	.25 (.03)	.52 (.04)	.61 (.04)

NOTE.—Power at level $\alpha = .05$ for regionwide test with use of 27 tSNPs when causal SNP is untyped, with estimated SE based on 200 simulated data sets; q and p are allele frequencies among nontransmitted alleles and transmitted alleles, respectively.

[a] In model 1, the causal SNP is VV (2, 134r) in block 4 with $q = 0.14$, which corresponds to marker 25 in the dense-SNP set.

[b] In model 2, the causal SNP is VV (24, 13r) in block 9 with $q = 0.25$, which does not correspond to any typed marker.

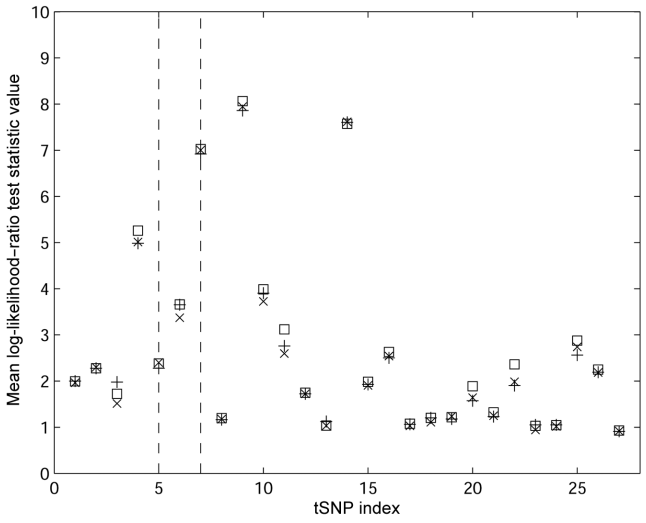

Figure 1. Mean log-likelihood–ratio test statistic value, across 200 simulated realizations, of single-point analysis (x), SBOV (+), and MOV (□). The horizontal axis represents the 27 tSNPs, with equal distances according to their map order. The causal SNP is VV (2, 134*r*) in block 4 (indicated by two vertical dashed lines from the 5th tSNP to the 7th tSNP) with $q = 0.14$ and $p = 0.30$.

Assessment of Goodness of Fit of the HMM for Background LD

With multipoint-likelihood methods for LD mapping, the issue arises of how to deal with background LD. An advantage of the single-block likelihood methods (SBOV and SBVV) over the full likelihood methods (MOV and MVV) is that, in the former, we are able to fully parametrize the LD model, so it is essentially nonparametric with respect to background LD, whereas, in the latter, we make some assumptions about the form of the background LD. On the other hand, the full likelihood method might be expected to make more-efficient use of the data and to lead to better localization. Therefore, it seems worth considering whether the background-LD model we use in the full likelihood method fits the data well. For our purposes, it is satisfactory to find a relatively parsimonious class of models that adequately captures the background-LD structure and that is computationally feasible for use in LD mapping. With sparse, highly informative markers such as microsatellites, background-LD models that assume independence or a Markov model of order 1 on markers have been used to model background LD, but such models may not be adequate to capture the structure of background LD in high-density SNP data.[22]

The model for background LD described in the "Full Multipoint-Association Analysis" section is an unconstrained Markov model of order 1, indexed by haplotype blocks rather than by markers, and we would like to assess the goodness of fit of this model to the nontransmitted haplotypes. Although we are interested in modeling only the nontransmitted haplotypes, because of the incomplete haplotype information, we are forced to model both the nontransmitted and the transmitted haplotypes. Our model for the transmitted haplotypes is also an unconstrained Markov model of order 1 indexed by haplotype blocks, with the connection between the parameterizations of the transmitted and nontransmitted haplotype models depending on the index d of the putative causal SNP (as detailed in the "Full Multipoint-Association Analysis" section). Because d is unknown, for the purposes of assessing goodness of fit, we parametrize the two Markov models independently. In other words, we simply assume separate and possibly different initial distributions and transition probabilities for the transmitted and nontransmitted haplotypes, where θ^T represents the parameters for the transmitted haplotypes and θ^N for the nontransmitted haplotypes.

A substantial difficulty arises, however, in assessment of significance for goodness of fit. It has been shown elsewhere[23] that, in a similar context, the ordinary parametric (or nonparametric, for that matter) bootstrap performs poorly for moderate-sized samples. The reason is that, for haplotypes that span several blocks, the multiblock haplotype-frequency distribution can be characterized as a multinomial distribution, with a large number of the outcomes having small nonzero probabilities. The difficulty with the ordinary parametric (or nonparametric) bootstrap is that the simulated realizations tend to have many fewer rare haplotypes than does the original data set (under the assumption that the simulated data set is of the same size as the original data set). Rare haplotypes in the original data set are commonly "lost" in the simulation, just by chance, and new haplotypes tend not to be created, because of the combination of model choice and estimation of the parameters by maximum likelihood. For example, if a particular haplotype in a block is not observed in data, then the corresponding parameter would have an MLE of 0, so the haplotype would not occur in any simulation.

To avoid this problem, we modify the model by introducing a smoothing parameter ϵ that reflects the fact that a new (previously unseen) haplotype is likely to be similar (only one site off, for instance) from a previously seen haplotype (see appendix B for details). We take ϵ to be a fixed constant and θ^T and θ^N to be unknown parameter vectors. When $\epsilon = 0$, we simply have the unconstrained Markov models for transmitted and nontransmit-

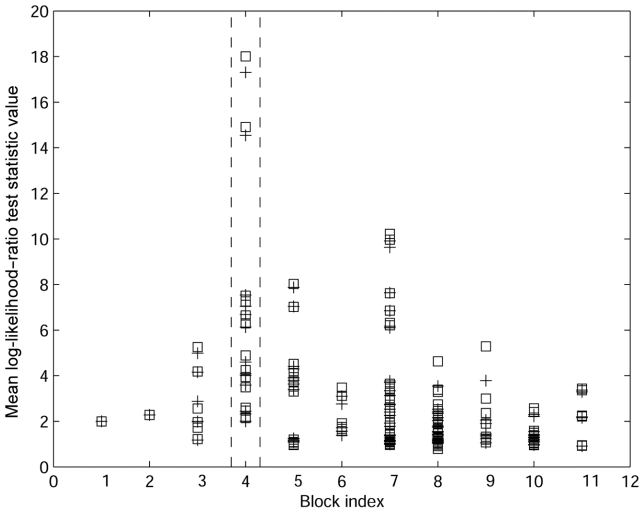

Figure 2. Mean log-likelihood–ratio test statistic value, across 200 simulated realizations, of SBVV (+) and MVV (□) for each VV in each block. The causal SNP is VV (2, 134*r*) in block 4 (indicated by two vertical dashed lines) with $q = 0.14$ and $p = 0.30$.

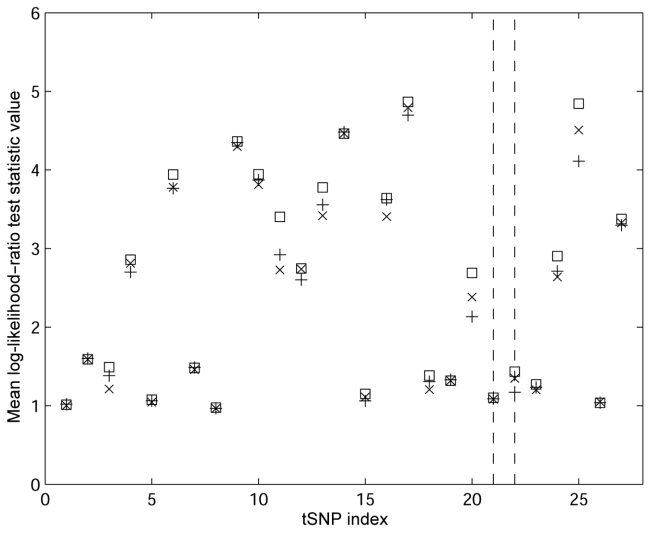

Figure 3. Mean log-likelihood–ratio test statistic value, across 200 simulated realizations, of single-point analysis (x), SBOV (+) and MOV (□). The horizontal axis represents the 27 tSNPs, with equal distances according to their map order. The causal SNP is VV (24, 13*r*) in block 9 (indicated by two vertical dashed lines from the 21st tSNP to the 22nd tSNP) with $q = 0.25$ and $p = 0.45$.

ted haplotypes described in the beginning of this subsection. When $\epsilon \neq 0$, the transmitted and nontransmitted haplotypes are no longer Markov but are instead hidden Markov. Note that, for small nonzero ϵ, the constrained model is a good approximation of the unconstrained Markov model, and it allows us to obtain the right type I error, for goodness of fit, with use of the parametric bootstrap.

To assess the significance of goodness of fit of our models to the data, we use a parametric bootstrap procedure described elsewhere.[23] We vary ϵ until we are able to obtain the correct type I error, which we assess by a nested set of simulations.[23] In the "Results" section, we apply this method to assess the goodness of fit of our background LD model compared with a CD data set.[2]

Assessment of Significance of Association

We consider tests of two different types of null hypothesis. (1) In a "site-specific" test, we test the null hypothesis that a particular variant is not associated with the trait. (2) In a "regionwide" test, we test the null hypothesis that an entire region is not associated with the trait. The test statistics T_{single}, T_{SBOV}, T_{SBVV}, T_{MOV}, and T_{MVV} are appropriate for site-specific tests, and we denote them generically by T_s, where *s* stands for "site-specific." To assess significance for a site-specific test, we can use either the χ^2_1 approximation for T_s under the null hypothesis or a simulation-based assessment of significance.

For a regionwide test, we specify one of T_{single}, T_{SBOV}, T_{SBVV}, T_{MOV}, and T_{MVV} to be T_s, and we use a test statistic of the form $T_m = \max T_s$, where the maximum is taken over all sites (or over all VVs, in the case of SBVV and MVV) in the region. To assess significance for T_m, we use the following procedure to simulate from the null distribution of T_m. Let Λ be the collection of inferred nontransmitted haplotypes in the original data, with each typed family contributing two haplotypes to Λ. (In practice, Λ can be obtained using software such as PHASE.[20]) We generate replicate trio data sets under the null hypothesis by sampling both transmitted and nontransmitted parental haplotypes, with replacement, from Λ. We then discard phase information to obtain the corresponding genotype data for each trio. Genotypes missing in the original data are also set to be missing in the simulated data. We simulate $\mu_1 = 1{,}000$ data sets, each containing *n* trios, where *n* is the number of trios in the original data set. Let t_m^i denote the observed value of T_m in the *i*th replicate data set, $1 \leq i \leq \mu_1$, and let t_m denote the observed value of T_m in the original data. An unbiased estimate of the *P* value associated with t_m is the proportion of $t_m^1, \ldots, t_m^{\mu_1}$ that are $\geq t_m$ (or we can add 1 to the numerator and denominator to ensure that the type I error is no larger than the nominal).

CD Data Set Used in Data Analysis and Simulation

We analyzed a previously published data set[2] consisting of 129 case-parents trios from a European-derived population who were genotyped for 103 common SNPs in a 500-kb region of 5q31 that may contain a genetic risk factor for CD. The whole region can be divided into 11 blocks of varying lengths (3–92 kb) and varying numbers of SNPs (4–32 SNPs per block). From the phased haplotype data (in which phase has been inferred from the trio data by use of an unpublished algorithm similar to PHASE[20]), generously provided by Mark Daly, we obtain the common haplotypes and other distinct rare haplotypes in each block. In the data, 10% of genotypes are missing, with a maximum missing per marker of 32% and a maximum missing per family of 49%. We omitted 12 families that have very low haplotype information in some blocks. The resulting data have 9.6% missing genotypes, with a maximum missing per marker of 28% and a maximum missing per family of 26%. Simulations are performed on the basis of this new data set, and subsequent use of the term "CD data set" refers to this new data set. Among these 103 dense SNPs, we choose 27 as tSNPs, where these distinguish the common haplotypes in each

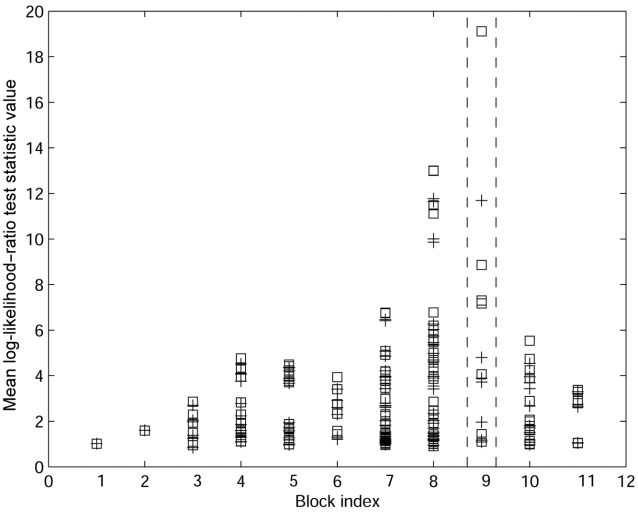

Figure 4. Mean log-likelihood–ratio test statistic value, across 200 simulated realizations, of SBVV (+) and MVV (□) for each VV in each block. The causal SNP is VV (24, 13*r*) in block 9 (indicated by two vertical dashed lines) with $q = 0.25$ and $p = 0.45$.

Table 3. Power of Regionwide Tests with Use of tSNPs When Causal SNP Is Typed

	Power of Method (SE)				
p	Single Point	SBOV	MOV	SBVV	MVV
.45	.86 (.02)	.88 (.02)	.87 (.02)	.78 (.03)	.80 (.03)
.40	.58 (.03)	.60 (.03)	.60 (.03)	.51 (.04)	.52 (.04)
.35	.21 (.03)	.21 (.03)	.22 (.03)	.15 (.03)	.18 (.03)

NOTE.—Power at level $\alpha = .05$ for regionwide test with use of 27 tSNPs when the causal SNP is typed (same SNP as in table 1), with SE estimated on the basis of 200 simulated data sets; $q = 0.275$ and p are allele frequency among nontransmitted alleles and transmitted alleles, respectively.

block. Where there are multiple equivalent choices of tSNPs for a block, we gave preference to those SNPs with fewer missing genotypes and to those sets of SNPs for which each haplotype determined by the tSNPs has low frequency unless it is one of the common haplotypes.

Assessment of Type I Error

To ensure that our proposed simulation procedure to assess significance of association works well, we used the following parametric bootstrap approach to assess its type I error. We simulated from the model described in appendix B, under the null hypothesis that $\theta^T = \theta^N = \theta$, for some choice of parameter θ and constant ϵ. To make our simulations relevant to our data example, we set $\epsilon = 0.001$ and then set θ to be $\tilde{\theta}$, the MLE of θ when the model is applied to the CD data set. (In the "Results" section, we demonstrate that this model does not show significant misfit to the data.) We simulate an "outer loop" of $\mu_2 = 200$ trio data sets, each of size $n = 117$, by simulating transmitted and nontransmitted haplotypes IID from the model. We then discarded the phase information to obtain the corresponding genotype data for each trio, including the same pattern of missing data, for each of the μ_2 replicate data sets. For each of the μ_2 data sets in the outer loop, we calculate a P value using the simulation procedure we have described above, which involves simulating $\mu_1 = 500$ "inner-loop" data sets for each outer-loop data set, for a total of $\mu_1\mu_2 = 10^5$ simulated data sets. Type I error is estimated as the proportion of times among the μ_2 outer-loop replicates that the P value is $\leqslant .05$.

Assessment of Power and Accuracy

We performed simulations to investigate the power and accuracy of localization of single-point and our various forms of multipoint-association analysis. Specifically, we address the following questions. (1) When the causal SNP is typed but with missing genotypes, how much extra power do we gain by using multipoint analysis instead of single point, and is use of the more computationally intensive full multipoint method a substantial improvement over just taking into account multipoint information within blocks? (2) When the causal SNP is untyped, do our multipoint methods that are specifically tailored to detect untyped variants perform better than the single-point and other multipoint methods, even though the former pay a greater price in terms of multiple comparisons? (3) Does multipoint-association analysis locate the causal SNP (typed or untyped) better than does single-point analysis? We propose three sets of simulations to answer these questions.

In the first set of simulations, we assess the power of the site-specific single-point, SBOV, and MOV association tests at a causal SNP that is typed with various proportions of missing genotypes. Let Λ denote the collection of inferred nontransmitted haplotypes in the CD data set.[2] Let q be the frequency in Λ of allele 1 at the causal SNP (taken to be the 5th tSNP, corresponding to the 31st dense SNP, which is in block 4). In the simulations, we vary the value of p, the frequency of allele 1 at the causal SNP among transmitted haplotypes. We draw nontransmitted haplotypes independently with replacement from Λ, and we draw transmitted haplotypes independently with replacement from Λ', with haplotype frequencies altered in the following way: if a haplotype having frequency f in Λ has allele 1 at the causal SNP, we set its haplotype frequency in Λ' to be $f' = fp/q$; otherwise we set its haplotype frequency in Λ' to be $f' = f(1-p)/(1-q)$. As a result, allele 1 at the causal SNP will have frequency p in Λ'. We obtain simulated genotype data by discarding the phase information and setting each genotype at the causal SNP to be missing independently, at random, with probability p_m. We perform this procedure both for the original "dense" set of 103 SNPs and for the "tag" set of 27 SNPs.

In the second set of simulations, we assume that the causal SNP is untyped and compare the power of the regionwide single-point, SBOV, MOV, SBVV, and MVV tests, to detect association on the basis of tSNPs. We also compare them in terms of their usefulness for localizing the causal variant by considering the mean log-likelihood–ratio test statistic for each variant in the region for each method. If the maximum of the mean test statistic is at or near the causal SNP for one method but not for another, we take this as an indication that the former method may be more useful for localization than the latter. The simulation is performed as above, with only the 27 tSNPs assumed to be typed and with the untyped causal variant taken to be either VV (2, 134r) in block 4, which corresponds to marker 25 in the dense-SNP set, or VV (24, 13r) in block 9, which does not correspond to any marker in the dense-SNP set.

In the last set of simulations, we compare the power of the regionwide single-point, SBOV, MOV, SBVV, and MVV association

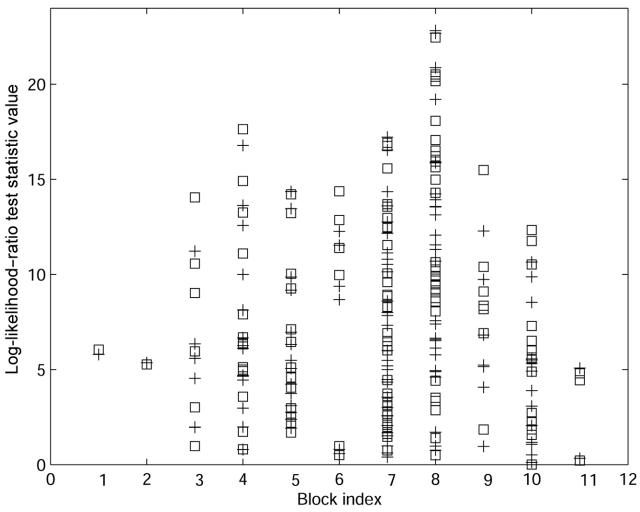

Figure 5. Log-likelihood–ratio test statistic values of SBVV (+) and MVV (□) for tSNPs of the CD data set.[2]

Table 4. Goodness of Fit of Background-LD Models for the CD Data Set[2]

	Goodness of Fit		
	LE[a] Model	Markov on Markers	Markov on Blocks with $\epsilon = .001$
Goodness-of-fit test statistic	LR_1[b]	LR_2[c]	Likelihood
Type I error (SE)	.05 (.007)	.06 (.008)	.05 (.02)
Goodness-of-fit *P* value (SE or CI[d])	0 (0–.003)	0 (0–.003)	.06 (.02)

NOTE.—To assess the type I error and goodness of fit of the background-LD models, μ_1 (number of inner-loop iterations) is set to be 1,000 for the LE model and the Markov model on markers and to be 200 for the Markov model on blocks with $\epsilon = .001$. μ_2 (number of outer-loop iterations) is set to be 1,000 for the LE model and Markov model on markers and to be 100 for the Markov model on blocks with $\epsilon = .001$.

[a] Markers are assumed to be independent.

[b] LR_1 is the ratio of the maximized likelihoods of the LE model and the Markov model on markers.

[c] LR_2 is the ratio of the maximized likelihoods of the Markov model on markers and the Markov model on haplotype blocks with $\epsilon = 0$.

[d] The ranges represent the 95% CI for the *P* value, based on an exact binomial distribution calculation.

tests in the situation in which the causal SNP is one of the typed tSNPs. The simulation is performed as above, with only the 27 tSNPs assumed to be typed and with the typed causal variant taken to be the same as in the first set of simulations.

Results

Simulation Results

We obtained the empirical type I error, at the nominal .05 level, for the site-specific tests that are based on single-point, SBOV, and MOV and for the regionwide tests based on single-point, SBOV, and SBVV. In each case, the empirical type I error is not significantly different from the nominal, which indicates that the empirical assessment of significance is working appropriately (data not shown).

Table 1 gives a power comparison of the site-specific single-point, SBOV, and MOV association tests based on dense SNPs or tSNPs, where the test is performed at a causal SNP typed with proportion p_m of genotypes missing. When $p_m = 0$, the single-point, SBOV, and MOV tests are identical, so their power to reject the null is the same. Not surprisingly, as p (frequency among transmitted) gets farther from q (frequency among nontransmitted), the power to detect association increases for all three methods. As p_m increases, the power of single-point analysis decreases the most, whereas, with dense SNPs, there is almost no change in the power of the multipoint methods (SBOV and MOV) when there are up to 30% missing genotypes. The multipoint methods retain power because they make use of the fact that nearby markers in the same block and/or neighboring blocks provide information on the marker that has missing genotypes. In the dense-SNP case, there are 10 other markers in the same block (block 4) with the causal SNP, which provide close-to-perfect information on the missing genotypes. In the tSNP case, there are only two other tSNPs in the same block with the causal SNP, so the power of MOV is higher than that of SBOV in that case, because MOV extracts extra information from other blocks.

Table 2 compares the power of the regionwide single-point, SBOV, MOV, SBVV, and MVV association tests that are based on tSNPs in the case where the causal SNP is untyped. It is clear that the two VV methods—SBVV and MVV—have much higher power than do the other methods, with MVV performing the best in all cases. Figures 1 and 2 compare the methods in terms of their ability to localize the untyped variant, for the case where the causal variant is in block 4. The VV methods—MVV and SBVV—are clearly far superior to the others in terms of their ability to localize the untyped variant. The results are even more dramatic in block 9, where there are more missing data (figs. 3 and 4), with the VV methods far outperforming

Table 5. Regionwide-Significant SNPs ($\alpha = .05$) in the CD Data Set[2]

				Test Statistic Value			
Marker	Block	q[a]	p_m[b]	Single Point	TDT	SBOV	MOV
26+	4	.36	.04	17.38*	15.87*	18.29*	18.25*
27+	4	.35	.09	16.13*	14.62*	18.31*	18.27*
28+	4	.35	.08	18.19*	16.13*	19.13*	19.09*
34+	4	.35	.09	13.89*	13.62*	16.06*	15.85*
39+	5	.38	.07	10.12	9.00	13.30*	13.13*
42	6	.38	.23	9.46	8.49	11.28*	13.73*
49+	7	.44	.16	8.15	8.34	11.33*	11.60
74+	7	.37	.07	9.46	9.14	18.95*	19.14*
78+	8	.45	.03	11.45*	10.98*	12.97*	13.09*
79	8	.14	.09	18.90*	16.89*	19.22*	19.68*
83	8	.45	.28	9.01	7.54	12.12*	9.55
92	10	.38	.29	9.78	9.45	15.29*	17.59*
93+	10	.38	.05	15.18*	14.37*	14.54*	14.47*

NOTE.—An asterisk (*) indicates SNPs showing regionwide significance ($\alpha = .05$) by at least one of the methods (single point, SBOV, and MOV). The regionwide thresholds for the single-point, TDT, SBOV, and MOV tests are 11.28, 10.66, 11.2, and 11.90, respectively, and are based on 1,000 simulations under the null hypothesis. Markers with a plus sign (+) were found elsewhere[7] to be associated with CD.

[a] q is allele frequency among nontransmitted alleles.

[b] p_m is the proportion of missing genotypes.

Table 6. Regionwide-Significant VVs Based on tSNPs of the CD Data Set[2]

Block and VV	Log-Likelihood–Ratio Test Statistic Value		Marker Index(es)[a]	tSNP[b]
	MVV	SBVV		
3:				
(1, 23*r*)	14.06	11.22	18	Yes
4:				
(1, 234*r*)	17.64	16.78	26, 27, 28, 34	Yes (34)
(12, 34*r*)	13.26	12.58	No	No
5:				
(1, 234*r*)	14.20	14.36	39	Yes
6:				
(1, 23*r*)	14.38	12.26	42	Yes
7:				
(1, 2345*r*)	16.92	16.54	74	Yes
(12, 345*r*)	15.58	10.54	49, 73	Yes (49)
(13, 245*r*)	12.58	12.16	No	No
(15, 234*r*)	13.56	13.60	No	No
(24, 135*r*)	13.70	10.80	No	No
8:				
(1, 2345*r*)	22.44	22.68	78	Yes
(2, 1345*r*)	18.08	15.92	79	No
(5, 1234*r*)	15.94	13.56	No	No
(12, 345*r*)	16.52	13.92	No	No
(13, 245*r*)	20.52	22.82	No	No
(14, 235*r*)	20.16	19.20	No	No
(15, 234*r*)	15.64	13.56	83	Yes
(23, 145*r*)	14.30	11.32	No	No
(24, 135*r*)	17.06	15.88	No	No
(25, 134*r*)	20.40	20.88	80	No
(35, 124*r*)	15.98	11.56	No	No
(45, 123*r*)	16.22	14.24	No	No
9:				
(1, 234*r*)	15.48	12.28	88, 91	No
10:				
(12, 34*r*)	12.34	10.66	97	No
(1, 234*r*)	11.76	8.54	92, 93, 95	Yes (95)

NOTE.—The thresholds of regionwide significance for SBVV and MVV are 11.48 and 11.24, respectively.

[a] Marker index(es) for VV if it is among the 103 typed markers in the CD data set[2]; "No" indicates it is not among the 103 typed markers.

[b] Indicates whether the VV corresponds to one of the tSNPs. A number in parentheses indicates which marker in the dense set was included in the tag set if the VV corresponds to more than one typed marker in the dense set.

the others. Note that the maximum log-likelihood ratios of the non-VV methods (single-point, SBOV, and MOV) tend to occur at SNPs with a relatively low proportion of missing genotypes. Furthermore, the non-VV methods are nearly equal when the proportion of missing genotypes is low. These two observations explain the fact that the non-VV methods have essentially equal power for the regionwide test with an untyped causal SNP (table 2).

Table 3 compares the power of the regionwide single-point, SBOV, MOV, SBVV, and MVV association tests based on 27 tSNPs when the causal SNP is typed (taken to be the same SNP as in the first set of simulations). Note that all five methods have lower power in table 3 than in table 1, because regionwide tests pay a penalty for multiple comparisons. Because there is a low proportion (2%) of missing genotypes at the causal SNP, the non-VV methods have almost identical power. In this context, the VV methods, as expected, have somewhat lower power than do the non-VV methods. This is because, with the VV methods, we test all combinations of common haplotypes in each block, so the penalty for multiple comparisons is higher. When the causal variant is untyped, the extra tests have a high payoff in terms of power and ability to localize (table 2 and figs. 1–4), whereas, when the causal variant is typed, they result in a slight loss of power. In the simulations with the causal variant typed, all five methods achieve their maximum mean log-likelihood–ratio values at the causal variant (figures not shown).

Application to CD Data Set

Goodness of fit of background LD models.—Table 4 verifies that our HMM with $\epsilon = 0.001$ is a not-too-unreasonable fit to the CD data ($P = .06$), with the type I error of our procedure verified to be at the nominal level. In comparison, models with an assumption of linkage equilibrium (LE) or a Markov model on markers are soundly rejected (table 4). We find that our parametric bootstrap procedure does not work (type I error = 1) for testing goodness of fit of the Markov model on blocks ($\epsilon = 0$ case). However, we note that the Markov model on blocks is approximated very accurately by the HMM with $\epsilon = 0.001$, with an average difference in log likelihoods of $<10^{-4}$ over 1,000 simulations. Therefore, because of its computational simplicity, we use the Markov model on blocks as the background-LD model in our full multipoint-likelihood methods (MOV and MVV). We see this as a reasonable compromise between overcomplicated and computationally intractable models on the one hand and simple models with obvious misfit on the other hand. Finally, we note that, because we use a nonparametric method to assess significance, model misfit would affect not the validity of the MOV and MVV tests but just their power. (The model is not used for the SBOV and SBVV tests.)

Table 7. Haplotype-Block Boundaries, with Different Methods, for the CD Data Set[2]

Method	No. of Blocks	Block Boundaries
PD[a]	11	8–9, 14–15, 24, 35, 40, 45, 76–77, 84–85, 91, 98
MDB[12b]	11	8, 14, 24, 36, 47, 57, 76, 86, 91, 98
htSNP[b]	10	8, 9, 11, 24, 40, 63, 80, 82, 91
DB[c]	5	24, 44, 77, 88
EQ10[d]	10	10, 20, 30, 40, 50, 60, 70, 80, 90
EQ20[e]	6	20, 40, 60, 80, 90

[a] Based on the pairwise D' value and the estimated recombination rates between adjacent sites.[2] The blocks are not strictly adjacent, and four markers (i.e., 9, 15, 77, and 85) do not belong to any block.

[b] Based on htSNP criterion[1] in the Haplblock[8] program.

[c] Based on haplotype diversity (see the Clayton Web site) in Haplblock[8] program.

[d] Every 10 consecutive markers are grouped as a block, and the last block has 13 markers.

[e] The first four blocks each have 20 markers, the fifth block has 10 markers, and the last block has 13 markers.

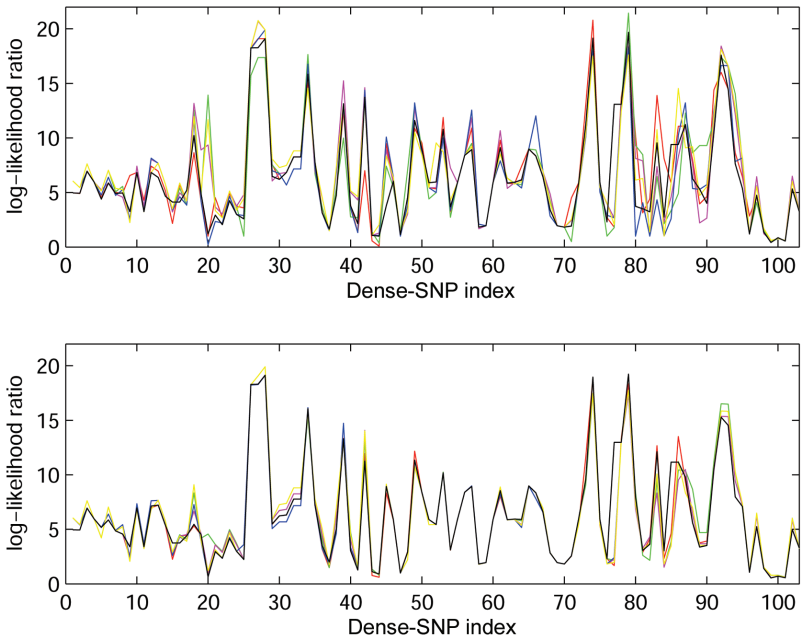

Figure 6. Log-likelihood–ratio test statistic values, from MOV (*top*) and SBOV (*bottom*), based on PD (*black line*), MDB (*red line*), htSNP (*green line*), DB (*pink line*), EQ10 (*blue line*), and EQ20 (*yellow line*) block boundaries for the CD data set.[2] In each graph, the horizontal axis represents the 103 SNPs, with equal distances according to the map order.

Analysis of the dense-SNP data.—All five methods detect significant association of CD to the region by use of regionwide tests that are based on the dense-SNP set. The estimated regionwide *P* value for the single-point method was 0 (95% CI 0–.003); for SBOV, .02 (95% CI 0–.006); for MOV, .02 (95% CI 0–.006); for SBVV, .01 (95% CI 0–.007); and for MVV, .02 (95% CI 0–.006). In every block, the most-significant VVs identified by the SBVV and MVV tests are in perfect LD with typed markers, so there is no evidence of untyped variation that would provide a stronger association signal. For that reason, we focus in this subsection on the results of the single-point, SBOV, and MOV analyses. The VV methods are expected to be most useful in the tSNP scenario considered in the "Analysis of the tSNP data" section.

Table 5 lists 13 SNPs significantly associated at the .05 level by at least one of the regionwide tests that are based on single-point, SBOV, or MOV, with TDT values included as a comparison. Notably, all 13 of the SNPs are regionwide significant at the .05 level on the basis of SBOV alone. Table 5 shows that the test-statistic values—and, hence, the significance of the associated SNPs—are much higher with the multipoint methods. It has been reported[7] that there are 11 SNPs in the same region that have alleles that are unique to a risk haplotype found to be significantly associated to CD by use of the TDT. Of those 11 SNPs, only 9 seem to be included in the CD data set, and all of these are included in table 5. The phased haplotype data are consistent with those nine SNPs, as well as with markers 42 and 92, each having an allele that is unique to the most common haplotype in each block. Of the 13 markers in table 5, 6 were detected only by the multipoint methods and not by the single-point or TDT methods. Marker 79 does not have an allele that is unique to the risk haplotype, but it is the most significant SNP by all four methods, which suggests it could be worthy of further investigation.

Analysis of the tSNP data.—We applied the multipoint mapping methods for VVs (SBVV and MVV) to the tSNPs of the CD data set. In our tSNP set, we include some but not all of the most significantly associated SNPs from the dense set, to see whether our MVV and SBVV methods can detect the presence of additional untyped variants associated with the trait. For example, there are eight markers with MOV log-likelihood–ratio test statistic value >14, whereas, in our set of tSNPs, we choose only two of them.

Figure 5 shows the log-likelihood–ratio test statistic values that are based on the SBVV and MVV methods for each VV in each block. Both MVV and SBVV achieve their maximums at VV (1, 2345*r*) in block 8, with MVV log-likelihood–ratio test statistic value of 22.44 and SBVV log-likelihood–ratio test statistic value of 22.68 (table 6), which again indicate that this region is significantly associated with the disease. (The thresholds of regionwide significance for SBVV and MVV are 11.48 and 11.24, respectively.) Table 6 lists VVs that are regionwide significant on the basis of MVV or SBVV and their relationship with the dense-SNP set and the tSNP set. For example, VV (2, 1345*r*) in block 8, which is regionwide significant on the basis of both MVV and SBVV, corresponds to marker 79 (which has the highest log-likelihood–ratio test statistic value

Table 8. Regionwide Significant SNPs, by MOV or SBOV with Different Block Boundaries, in the CD Data Set[2]

Significant Marker	Block Algorithm											
	PD		MDB		htSNP		DB		EQ10		EQ20	
	M[a]	S[b]	M	S	M	S	M	S	M	S	M	S
18					*		*				*	
20							*					
26 +	*	*	*	*	*	*	*	*	*	*	*	*
27 +	*	*	*	*	*	*	*	*	*	*	*	*
28 +	*	*	*	*	*	*	*	*	*	*	*	*
34 +	*	*	*	*	*	*	*	*	*	*	*	*
39 +	*	*	*	*	*	*	*	*	*	*	*	*
42	*	*		*	*	*	*	*	*	*	*	*
49 +		*		*	*		*		*			
57									*			
66									*			
73			*	*								
74 +	*	*	*	*	*	*	*	*	*	*	*	*
78 +	*	*	*	*	*	*	*	*	*	*	*	*
79	*	*	*	*	*	*	*	*	*	*	*	*
83		*	*									
86				*							*	
87							*		*			
91			*						*		*	
92	*	*	*	*	*	*	*	*	*	*	*	*
93 +	*	*	*	*	*	*	*	*	*	*	*	*
94							*				*	

NOTE.—Regionwide significant SNPs are indicated by an asterisk (*). Markers with a plus sign (+) were found elsewhere to be associated with CD.[7]

[a] Significant with MOV.

[b] Significant with SBOV.

based on the MOV analysis) in the dense-SNP set, but it does not correspond to any of the tSNPs, which suggests that our MVV and SBVV methods have power to detect such "untyped" variants that are strongly associated with the trait. Furthermore, we notice that, from block 3 to block 10, VVs that correspond to common haplotype 1—namely, (1, 23*r*) in blocks 3, 6, and 10; (1, 234*r*) in blocks 4, 5, and 9; and (1, 2345*r*) in blocks 7 and 8—are all found to be significant, which suggests that the first common haplotype across the whole region (or, more precisely, from block 3 to block 10) is strongly associated with the trait. This observation is consistent with the risk haplotype found in previous studies.[7]

SBOV and MOV with different block boundaries.—In our multipoint mapping methods, we assume that haplotype-block boundaries have been identified. There are many ways to define a haplotype block, and we are interested in the question of whether our multipoint mapping method is robust to different haplotype-block definitions. To answer this question, we considered six different haplotype-block definitions: PD (pairwise D')[2] is based on the pairwise D' value and the estimated recombination rates between adjacent sites, MDB (MDBlocks)[12] is based on the minimum-description-length principle, htSNP[1] aims to minimize the total number of SNPs needed to represent the common haplotypes in the whole region, and DB (diversity-based) is based on a haplotype diversity measure (see the Clayton Web site). As a complement to these more principled block-boundary methods, we present two simplistic methods, EQ10 and EQ20, which simply put block boundaries every ~10 SNPs and every ~20 SNPs, respectively. In particular, with 103 SNPs in the region of interest, EQ10 puts 10 SNPs in each of the first nine blocks and 13 SNPs in the 10th block, whereas EQ20 puts 20 SNPs in each of the first four bocks, 10 SNPs in the 5th block, and 13 SNPs in the 6th block. (We tried combining the 5th and 6th blocks, but the resulting high level of haplotype diversity resulted in slower computations.) Table 7 lists the block boundaries from those six methods. The block boundaries resulting from the application of the PD, MDB, htSNP, and DB methods to the CD data set[2] have been reported elsewhere.[2,12]

We apply SBOV and MOV to the CD data set with these different block boundaries. Figure 6 shows the log-likelihood–ratio test-statistic curves of MOV and SBOV, based on the block boundaries of PD, MDB, htSNP, DB, EQ10, and EQ20, in different colored lines. The six different log-likelihood–ratio test-statistic curves are remarkably close to each other, suggesting that the methods are reasonably robust to choice of block boundaries. Table 8 lists the SNPs that are regionwide significant by MOV and SBOV with use of different block boundaries. Of the nine SNPs found elsewhere to be associated with CD,[7] all but one (SNP 49) are found to be significant by all 12 tests, regardless of

choice of block boundaries. For marker 49, in the cases in which it is not regionwide significant, its MOV value or SBVV value is close to the significance threshold.

In this example, the multipoint mapping methods (SBOV and MOV) seem rather robust to different block boundaries, in the sense that the same SNPs tend to be detected with different choices of block boundaries, and the log-likelihood–ratio test-statistic curves have similar shape as well.

Discussion

We have developed novel methods, for multipoint LD mapping of binary traits, that make explicit use of haplotype-block structure. In particular, our SBVV method has a desirable combination of high power, accurate localization, relatively few modeling assumptions, and computational feasibility.

Multipoint LD mapping has major advantages over single-point analysis when a variant of interest is untyped, in which case genotypes at multiple markers can often jointly provide considerably more information on the untyped variant than single markers can. One difficulty in using this information is the large number of possible multimarker tests, which can reduce power when appropriate correction is made for multiple comparisons. Our strategy is to make explicit use of inferred block structure to determine a relatively small number of VVs to test. In our simulations, this strategy (used in MVV and SBVV methods) has much higher power than does single-point analysis and is vastly more accurate for localization.

Another difficulty with multipoint methods is that they are computationally more challenging than single-point methods and may rely on modeling assumptions about background LD. We avoid both of these difficulties in our SBVV method by using the inferred block structure to approximate the full multipoint likelihood by the local multipoint likelihood for the block. With a single 3.4-GHz Pentium 4 processor with 1 GB memory, it took 100 s to analyze a 500-kb region with 27 tSNPs and 138 VVs, where this includes computing the SBVV statistic for each VV and obtaining both nominal and regionwide *P* values based on 1,000 simulated replicates. To get a rough idea of how this would scale up, we estimate that the time to analyze 250,000 tSNPs genomewide by this method, assuming an average of four tSNPs per block and four common haplotypes per block—resulting in ~875,000 VVs and including 1,000 simulated replicates to obtain genomewide as well as nominal *P* values—would be ~7 d (the time is approximately linear in the number of VVs). Our code is not optimized, so future speed-ups may be substantial. Note that the 1,000 replicates would be able to be trivially parallelized with multiple processors.

In our multipoint LD-mapping methods, we assumed that haplotype blocks and their common haplotypes have been inferred. Comparison of results on the basis of six different methods for inferring blocks indicates that our methods are reasonably robust to the choice of block boundaries. Even grouping every 10 or 20 consecutive markers in a block was effective and might serve well for a first-pass analysis of data.

Our methods are designed to detect common SNPs (typed or untyped) that are associated with disease. If there is a small number of relatively common causal SNPs in a block, then the SBVV and MVV methods would be expected to detect them, because they consider all possible partitions of the set of common haplotypes into two subsets. Similarly, in the case of multiple rare untyped variants, if the rare variants, as a group, are associated with a subset of the haplotypes that can be identified by the tSNPs, then the SBVV and MVV should be able to detect the association. However, if there is no subset of tagged haplotypes that happens to be associated with the set of untyped rare variants, then one would expect that association would be difficult or impossible to detect.

We have approached the problem of using fine-scale LD information in mapping by summarizing that information in terms of inferred haplotype blocks and common haplotypes. We then used this information both to make the computations faster and to limit the number of multimarker tests. Another strategy that would target specifically those SNPs typed in HapMap but not in the mapping study would be to use the HapMap information to determine specific haplotypes or haplotype combinations that approximately query these SNPs.

Acknowledgments

We are grateful to Mark Daly for providing us with phased haplotype data and to two anonymous referees for helpful comments. This work is supported by National Institutes of Health grants HG001645 and HL084715.

Appendix A

When genotype data are available for d (the SNP of interest) for all individuals, we have $T_{\text{SBOV}} = T_{\text{single}}$. This property is a consequence of the fact that the parameter (p,q) is specified independent of the background LD parameter $\alpha_d = (\alpha_{h,d}, h \in S)$. Consider a single trio with complete genotype data available at SNP d but with some genotype data possibly missing at other SNPs in the block. The likelihood for this trio, under the model, can be written $L(p,q,\alpha_d) = L_1(p,q)L_2(\alpha_d)$, where $L_1(p,q) = p^{n_t}(1-p)^{2-n_t}q^{n_u}(1-q)^{2-n_u}$ and

$$L_2(\alpha_d) = \sum_{(h_1,h_2,h_3,h_4)\in S^4} \left(\prod_{k=1}^{4} \alpha_{h_k,d}\right) P[Y \mid \left(X^{MT},X^{MN},X^{FT},X^{FN}\right) = (h_1,h_2,h_3,h_4)] \ ,$$

where X and Y are as defined in the "Full Multipoint-Association Analysis" section. Thus, the likelihood factors into two independently parameterized parts, and this factorization continues to hold when the likelihoods for independent trios are multiplied together. Since $L_2(\alpha_d)$ does not change between the null and alternative models and α_d is specified independent of p and q, the maximized $L_2(\alpha_d)$ is the same in both the null and alternative models and so cancels out of the likelihood ratio. Note that $L_1(p,q)$ is just the single-point likelihood. When there are incomplete genotype data for the SNP of interest, this factorization no longer holds (because L_1 moves inside the summation over the complete data), and the two likelihood-ratio test statistics T_{SBOV} and T_{single} are, in general, different. These arguments easily generalize to T_{MOV}.

Appendix B

In assessing goodness of fit of the HMM for background LD, we consider a smoothed version of the model. First, consider the smoothed single-block model

$$P(X = i) = (1 - K\,\epsilon)f_i + \epsilon \sum_{i':\,|i'-i|=1} f_{i'} \tag{B1}$$

for $i \in S$. Here, K is the number of SNPs in the block, ϵ is a fixed constant with $0 \leqslant \epsilon \leqslant 1/K$, $(f_i, i \in S)$ is a parameter vector, and the summation is over all haplotypes i' that differ from i at exactly one site. Equation (B1) can be interpreted as saying that, to construct a realization of X, a haplotype is drawn from f, and, then, with probability $1 - K\epsilon$, the observed haplotype is the same as the one drawn, whereas, with probability $K\epsilon$, it is one site off from the one drawn, where the site that differs is drawn uniformly from the typed SNPs in the block. An estimate of f can be obtained by maximum-likelihood estimation.

We extend this approach to multiblock haplotypes for trios by applying a similar smoothing idea to our Markov model. Let $X_t = (X_t^{MT},X_t^{MN},X_t^{FT},X_t^{FN})$ for $1 \leqslant t \leqslant B$ have the same meaning as in the "Full Multipoint-Association Analysis" section. Let $h = (h^{MT},h^{MN},h^{FT},h^{FN}) \in S^4$ be a quadruple of possible haplotypes and $h_t = (h_t^{MT},h_t^{MN},h_t^{FT},h_t^{FN})$ be the quadruple of haplotypes for block t obtained by restricting h to block t. Then, our smoothed model is

$$P[(X_1,\ldots,X_B) = (x_1,\ldots,x_B)] = \sum_{h\in S^4} f(h^{MT},\theta^T)f(h^{MN},\theta^N)f(h^{FT},\theta^T)f(h^{FN},\theta^N)\prod_{t=1}^{B} G_t(x_t,h_t) \ ,$$

where $f(h,\theta^T)$ and $f(h,\theta^N)$ represent, respectively, the probability of haplotype h under the unconstrained Markov models (indexed by block) for transmitted and nontransmitted haplotypes. Here, $G_t(x_t,h_t) = (1 - 4K_t\epsilon)I_{x_t=h_t} + \epsilon I_{|x_t-h_t|=1}$, where $|x_t - h_t| = 1$ denotes the event that the quadruple of haplotypes x_t has exactly three of its haplotypes equal to the corresponding haplotypes of h_t and one of its haplotypes differing from the corresponding haplotype of h_t by exactly one site. This form for G_t is obtained by applying equation (B1) to the four independent chains and neglecting higher-order terms in ϵ.

Web Resources

The URLs for data presented herein are as follows:

D. Clayton, http://www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf (for "Choose a set of haplotype tagging SNPs from a large set of diallelic loci")

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for CD)

References

1. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, et

al (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723
2. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232
3. Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222
4. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229
5. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
6. Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597
7. Rioux JD, Daly MJ, Silverberg MS, Lindblad K, Steinhart H, Cohen Z, Delmonte T, Kocher K, Miller K, Guschwan S, et al (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228
8. Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002) A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA 99:7335–7339
9. Johnson GCL, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, et al (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237
10. Zhang K, Calabrese P, Nordborg M, Sun F (2002) Haplotype block structure and its applications to association studies: power and study design. Am J Hum Genet 71:1386–1394
11. Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 14:908–916
12. Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73:336–354
13. Lin Z, Altman RB (2004) Finding haplotype tagging SNPs by use of principal components analysis. Am J Hum Genet 75: 850–861
14. Carlson CS, Eberle MA, Rieder MJ, 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
15. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D (2005) Efficiency and power in genetic association studies. Nat Genet 37:1217–1223
16. Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MR (2005) An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet 37:1320–1322
17. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516
18. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38
19. Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, et al (2002) Association of the *ADAM33* gene with asthma and bronchial hyperresponsiveness. Nature 418:426–430
20. Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989
21. Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3:1–8
22. Strahs A, McPeek MS (2003) Multipoint fine-scale linkage disequilibrium mapping: importance of modeling background LD. In: Goldstein DR (ed) Science and statistics: a Festschrift for Terry Speed. Institute of Mathematical Statistics Lecture Notes Monograph Series. Vol 40. Institute of Mathematical Statistics, Beachwood, OH, pp 343–366
23. Zheng M, McPeek MS (2004) Parametric bootstrap for assessment of goodness of fit of models for block haplotype structure. In: Istrail S, Waterman MS, Clark AG (eds) Springer lecture notes in computer science series: computational methods for SNPs and haplotype inference. Vol 2983. Springer Verlag, Berlin, pp 113–123