
Bayesian Spatial Modeling of Haplotype Associations

Duncan C. Thomas Daniel O. Stram David Conti John Molitor
Paul Marjoram

University of Southern California, Los Angeles, Calif., USA

Key Words

Candidate gene associations · Haplotypes · Single nucleotide polymorphisms · Bayesian methods · Spatial statistics · Linkage disequilibrium mapping · Case-control studies

Abstract

We review methods for relating the risk of disease to a collection of single nucleotide polymorphisms (SNPs) within a small region. Association studies using case-control designs with unrelated individuals could be used either to test for a direct effect of a candidate gene and characterize the responsible variant(s), or to fine map an unknown gene by exploiting the pattern of linkage disequilibrium (LD). We consider a flexible class of logistic penetrance models based on haplotypes and compare them with an alternative formulation based on unphased multilocus genotypes. The likelihood for haplotype-based models requires summation over all possible haplotype assignments consistent with the observed genotype data, and can be fitted using either Expectation-Maximization (E-M) or Markov chain Monte Carlo (MCMC) methods. Subtleties involving ascertainment correction for case-control studies are discussed. There has been great interest in methods for LD mapping based on the coalescent or ancestral recombination

graphs as well as methods based on haplotype sharing, both of which we review briefly. Because of their computational complexity, we propose some alternative empirical modeling approaches using techniques borrowed from the Bayesian spatial statistics literature. Here, space is interpreted in terms of a distance metric describing the similarity of any pair of haplotypes to each other, and hence their presumed common ancestry. Specifically, we discuss the conditional autoregressive model and two spatial clustering models: Potts and Voronoi. We conclude with a discussion of the implications of these methods for modeling cryptic relatedness, haplotype blocks, and haplotype tagging SNPs, and suggest a Bayesian framework for the HapMap project.

Copyright © 2003 S. Karger AG, Basel

Introduction

Because of their high density throughout the human genome (about one every 200 base pairs [1]), relative ease of high-throughput genotyping, and rapidly growing catalogs of variants [2], single nucleotide polymorphisms (SNPs) are ideally suited for testing candidate gene associations and for linkage disequilibrium (LD) mapping of unknown genes. Furthermore, the recent recognition that LD tends to be concentrated in blocks of limited haplo-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2003 S. Karger AG, Basel
0001-5652/03/0563-0032\$19.50/0

Accessible online at:
www.karger.com/hhe

Duncan O. Thomas
Department of Preventive Medicine, University of Southern California
1540 Alcazar St, CHP-220
Los Angeles, CA 90089-9011 (USA)
Tel. +1 323 442 1218, Fax +1 323 442 2349, E-Mail dthomas@usc.edu

type diversity has sparked great interest in the use of haplotypes for both purposes. We review here a broad range of methods that have been proposed for haplotype association studies, with particular emphasis on Bayesian methods. We begin, however, with a general likelihood-based framework for inference on haplotype associations in case-control studies using unrelated individuals.

Genotype-Based and Haplotype-Based Penetrance Models

Let Y_i denote the phenotypes of a sample of $i = 1, \dots, n$ subjects. For most purposes, we will restrict attention to case-control data, where $Y = 1$ indicates a case and $Y = 0$ a control, although the general linear model we propose can be applied to most any phenotype by appropriate selection of a link function. Suppose we also observe for each subject a vector of $l = 1, \dots, L$ SNP genotypes $\mathbf{G}_i = (G_{il})$, where $G_{il} = 0, 1, 2$ indicates the number of copies of a particular allele (conventionally, the rarer allele). Using standard logistic regression approaches, we might consider a penetrance model of the form

$$\text{logit Pr}(Y_i = 1 | \mathbf{G}_i) = \beta_0 + \sum_l \beta_l Z(G_{il}) + \dots \quad (1)$$

where $Z(G)$ denotes some coding of the genotypes incorporating any assumptions about dominance and ‘...’ indicates the possibility of including covariates or additional locus-by-locus or gene-environment interaction terms. We call this a genotype-based model and note that such an approach does not require knowledge of the genotypes’ phase, i.e., how they are arranged into haplotypes. We contrast this approach with a haplotype-based model of the form

$$\text{logit Pr}(Y_i = 1 | \mathbf{H}_i) = \gamma_0 + \gamma_{h_{i1}} + \gamma_{h_{i2}} + \dots \quad (2)$$

where $\mathbf{H}_i = (h_{i1}, h_{i2})$ designates a pair of haplotypes h in the space H of all haplotypes represented in the population. Since these haplotypes can usually not be determined with certainty from the observed genotypes, the full ‘prospective’ likelihood is given by

$$L_{(P)}(\gamma, \mathbf{q}) = \Pr(\mathbf{Y}, \mathbf{G}) = \prod_{i=1}^n \sum_{\mathbf{h} \sim \mathbf{G}_i} \Pr(Y_i | \mathbf{H}_i = \mathbf{h}; \gamma) \Pr(\mathbf{H}_i = \mathbf{h} | \mathbf{q}) \quad (3)$$

where the summation is over the set $\mathbf{h} \sim \mathbf{G}_i$ of haplotype pairs that are compatible with each individual’s observed genotypes (where genotypes are missing, the summation would be over all haplotypes that are compatible with the available genotype data). In addition to the haplotype rel-

ative risks γ , this likelihood is also a function of the population haplotype frequencies $\mathbf{q} = (q_h)_{h \in H}$; assuming the population is in Hardy-Weinberg (H-W) equilibrium, $\Pr(\mathbf{H} | \mathbf{q}) = q_{h_1} q_{h_2}$.

We can fit the model using the E-M algorithm [3–6], provided the number of loci (or haplotypes) is not too large. This is commonly done using a two-stage procedure, in which \mathbf{q} is first estimated from the controls (or in the total sample ignoring the phenotypes) and then treated as fixed in a score test of $H_0: \gamma = 0$ based on likelihood (3). The justification of combining cases and controls in this test is that under H_0 , there is no difference in haplotype frequencies between the two groups, but this is of course violated if the aim is estimation of γ rather than hypothesis testing. Stram et al. [7] have extended this basic E-M approach to a single-stage joint estimation of γ and \mathbf{q} . Their E-step entails calculation of the expectation of the number of copies $N_h(\mathbf{G}_i)$ of each haplotype $h \in H$ given both \mathbf{G}_i and Y_i , conditional on the current estimates of the parameters, and the M-step entails maximization of the complete-data likelihood for γ and \mathbf{q} using these expectations.

This calculation can become unwieldy if there are a large number of possible haplotypes, although in practice the number that are actually observed in human populations over reasonably small regions tends to be modest, even if many SNPs are included [8, 9], and in any event the analysis could always be restricted to the subset of those with some minimal estimated population frequency. (We will revisit this issue when we consider haplotype block structure below.) No such restriction is needed, however, if Markov chain Monte Carlo (MCMC) methods are used [10–12]. Here, the basic idea is to alternate between sampling from $[\mathbf{H}_i | \mathbf{G}_i, Y_i; \gamma, \mathbf{q}]$, $[\gamma | \{Y_i, \mathbf{H}_i\}]$, and $[\mathbf{q} | \{\mathbf{H}_i\}]$, where $[\cdot | \cdot]$ denotes the respective full conditional distributions. The first of these updates can easily be done by a Metropolis-Hastings step, proposing a switch of the alleles at a single locus or a contiguous segment. The remaining updates use standard MCMC moves. We have implemented similar MCMC approaches to haplotype assignment in pedigrees, where a change to any individual’s haplotype must be propagated through the rest of the pedigree and the Hastings ratio is now also a function of the recombination fractions. Yet another approach uses estimating equations methods [13].

A number of authors have considered the relative efficiency of genotype-based and haplotype-based models [5, 14–18]. Because LD between individual SNPs is typically too high to fit multiple logistic models involving all SNPs simultaneously, most of these authors have compared a

haplotype-based model (typically a multiple degree of freedom test) against the best-fitting single-SNP models with Bonferroni correction for multiple comparisons. Of course, it is also necessary to sum over the unknown phases in any haplotype-based test. Results of such comparisons have been variable, depending upon the specific tests compared and what is assumed about the extent of LD and the causal model for penetrance. If LD is high and a single haplotype contains the causal variant, a haplotype-based test can be more powerful, despite the larger degrees of freedom and the uncertain phases. On the other hand, if LD is low (so that phase is difficult to infer) or if there is a single causal variant among the set of available SNPs and its effect is spread out over a number of different haplotypes, then a genotype-based test can be more powerful.

Conti and Witte [19] have proposed a hierarchical genotype-based model, in which each β_l is estimated univariately in the first-level model, but a second-level regression model of the form $\boldsymbol{\beta} \sim N(\mathbf{X}'\boldsymbol{\alpha}, \Sigma(x))$ is used to smooth the $\hat{\beta}$ s, where \mathbf{X} is a matrix of ‘prior covariates’ for each SNP and $\Sigma(x)$ is a covariance matrix that could depend upon the location x of the postulated causal variant. For example, the prior covariates could include such characteristics as the putative functional significance, the location of a variant (e.g. which haplotype block it is located in), or which common haplotype(s) contain the variant. We are currently exploring extensions of this approach to incorporate SNP \times SNP interactions in the hopes of capturing more of the information in haplotypes without actually having to resolve phases. Bayesian model selection or model averaging techniques [20, 21] may prove useful here to allow for our uncertainty about whether particular SNPs or interaction terms should be included in the model (see also Conti et al. [22, this issue] for an application of Bayes model averaging in a different context).

Likelihoods and Case-Control Ascertainment

For logistic models with fully observed covariates, such as the genotype model (eq. (1)), there is a well known equivalence between the ‘prospective likelihood’ $\Pr(\mathbf{Y}|\mathbf{G}, n(\mathbf{Y}))$ (where $n(\mathbf{Y})$ denotes the number of subjects with $Y = 1$) and the ‘retrospective likelihood’ $\Pr(\mathbf{G}|\mathbf{Y})$ [23, 24]. Proof of this relationship is complex and relies on two key observations: first, that the odds ratio for $[\mathbf{Y}|\mathbf{G}]$ is the same as the odds ratio for $[\mathbf{G}|\mathbf{Y}]$; and second, that consistent estimates of the parameters of this odds ratio can be

obtained by allowing the distribution of \mathbf{G} to be unspecified or to depend on parameters that are independent of the odds ratio. For further discussion in the context of family studies see [25, 26]. This is the fundamental justification for the appropriateness of the prospective likelihood for case-control studies, even though it is Y that is sampled and G that is observed.

For models with incompletely observed covariates, such as the haplotype model (eq. (2)), this equivalence no longer holds. The likelihood given in eq. (3) is appropriate for cohort studies, but not for case-control studies without additional correction for the differential sampling fractions of cases and controls. Stram et al. [7] have shown that naive application of the cohort likelihood to case-control data can lead to substantial bias in the population haplotype frequency estimates, essentially because the high risk haplotypes are overrepresented in a case-control sample. Because the estimates of $\boldsymbol{\gamma}$ and \mathbf{q} are not independent, this can also lead to some bias in $\hat{\boldsymbol{\gamma}}$, although the magnitude of this bias is generally small and depends upon how accurately the haplotypes can be predicted from the genotypes. This bias can be eliminated in three ways. The standard conditional likelihood, $\Pr(\mathbf{Y}|\mathbf{G}, n(\mathbf{Y}))$, can be computationally daunting for large strata because of the need to sum over all possible permutations of \mathbf{Y} with the same $n(\mathbf{Y})$ as that observed [27]. A simpler approach, but still prospective, is to condition on the marginal probability of each subject being ascertained,

$$L_{(PA)}(\boldsymbol{\gamma}, \mathbf{q}) = \Pr(\mathbf{Y}|\mathbf{G}, Asc) = \prod_{i=1}^n \frac{\sum_{\mathbf{h} \sim G_i} \pi_{Y_i} \Pr(Y_i|\mathbf{G}_i) \Pr(\mathbf{H}_i = \mathbf{h})}{\sum_{\mathbf{h} \sim G_i} [\pi_1 \Pr(Y_i = 1|\mathbf{H}_i = \mathbf{h}) + \pi_0 \Pr(Y_i = 0|\mathbf{H}_i = \mathbf{h})] \Pr(\mathbf{H}_i = \mathbf{h})}$$

where π_1 and π_0 are the case and control sampling fractions respectively. This likelihood requires knowledge of the sampling fractions π , which are readily available for a nested case-control study within a cohort, but would require knowledge of the marginal disease rate in the population for a conventional population-based case-control study. Alternatively, the retrospective likelihood

$$L_{(R)}(\boldsymbol{\gamma}, \mathbf{q}) = \Pr(\mathbf{G}_i|Y_i) = \frac{\sum_{\mathbf{h} \sim G_i} \Pr(Y_i|\mathbf{H}_i = \mathbf{h}) \Pr(\mathbf{H}_i = \mathbf{h})}{\sum_{\mathbf{h} \in H} \Pr(Y_i|\mathbf{H}_i = \mathbf{h}) \Pr(\mathbf{H}_i = \mathbf{h})} \quad (4)$$

can be used. Epstein and Satten [28] show that this likelihood can be expressed in a simpler manner if parameterized in terms of the haplotype frequencies *in controls* and the odds ratios; this also avoids having to assume the haplotypes are in H-W equilibrium. Kraft and Thomas

[29] compare the performance of the retrospective and (ascertainment-corrected) prospective likelihood approaches in the case of fully observed covariates and conclude that the latter often provides a more efficient estimator of β , despite the need to address the additional \mathbf{q} parameters. Whether this is true in the case of incompletely observed covariates remains to be studied.

Multiple Comparisons and Bayesian Smoothing or Clustering

If there are many haplotypes in H , we are confronted with the related problems of multiple comparisons and sparse data, for which Bayesian shrinkage estimators offer a natural solution. In doing so, we wish to exploit the notion that structurally similar haplotypes in the neighborhood of a disease predisposing locus are more likely to harbor the same susceptibility allele and hence to have similar γ s. This problem is very similar to that considered in the Bayesian spatial smoothing and spatial clustering literature, with diverse applications such as spatial modeling of disease rates in environmental epidemiology [30–33]. Thomas et al. [34] and Molitor et al. [35] considered a conditional autoregressive (CAR) model of the form

$$\gamma \sim N(\mathbf{0}, \sigma^2 \mathbf{I} + \tau^2 \mathbf{W})$$

where \mathbf{W} is a matrix of ‘similarities’ of each pair (h, k) of haplotypes, such as the length $L_{hk}(x)$ of the segment shared identical by state (IBS) surrounding a candidate mutation location x . Standard MCMC methods are used to update γ , σ^2 , and τ^2 . The location x is updated by a Metropolis-Hastings move, proposing either a small random walk in the neighborhood of the current location or an entirely new location anywhere in the region.

More recently, we have been considering the Potts [36] and Voronoi [37] spatial clustering models, of the form

$$\text{logit Pr}(Y_i = 1 | \mathbf{H}_i) = \delta_0 + \delta_{c_{h1}} + \delta_{c_{h2}}$$

where c_h denotes the ‘cluster’ to which haplotype h belongs and δ_c is a relative risk parameter common to all haplotypes assigned to a particular cluster c . (As a further generalization, one could allow $\gamma_h \sim N(\delta_{c_h}, \sigma^2)$). For the Potts model,

$$\Pr(\mathbf{c}) = \frac{\exp[\psi \sum_{hk} W_{hk}(x) I(c_h = c_k)]}{\Psi[C, \psi, \mathbf{W}(x)]}$$

where $\Psi[C, \psi, \mathbf{W}(x)]$ is a normalizing constant equal to the sum of the numerator over all possible partitions. This leads to a simple expression for the full conditional distri-

bution of the cluster assignment for any particular haplotype,

$$\Pr(c_h = c | \mathbf{c}_{-h}) = \frac{\exp[\psi \sum_{hk} W_{hk}(x) I(c_k = c)]}{\sum_{c'=1}^C \exp[\psi \sum_{hk} W_{hk}(x) I(c_k = c')]}$$

but updates of the parameters C , ψ , and x involve the normalizing constant Ψ , which can be quite complex. In the Voronoi model, haplotypes are assigned deterministically to the cluster containing the ‘nearest’ ancestral haplotype A_c [38], thereby avoiding the need to consider any normalizing constants. As in the CAR model, MCMC methods are used to update \mathbf{c} , ψ , x , and A_c . Reversible jump MCMC methods [39] are needed to update the number of clusters C , but these calculations are greatly simplified by integrating out the γ , as described by Denison and Holmes [33].

One can combine the Potts and Voronoi approaches by using the Voronoi model, but assigning haplotypes to centers probabilistically using a Potts model approach. Specifically, one can express the probability of haplotype assignment to a cluster c given the cluster center A_c as

$$\Pr(c_h = c | \mathbf{c}_{-h}) = \frac{\exp[\psi \sum_h W_h(x, A_c)]}{\sum_{c'=1}^C \exp[\psi W_h(x, A_{c'})]}$$

Here the normalizing constant is simply a sum over the number of clusters, not the sum over all possible haplotype allocations. By incorporating an extra step of estimating latent cluster centers, we can reduce the dimensionality of the parameter space.

Relationship to Coalescent Methods

We view the spatial smoothing and spatial clustering approaches as a relatively simple ‘empirical’ approximation to the more formal coalescent methods that have received a great deal of attention as a possible approach to LD mapping [40–44]. The coalescent, introduced by Kingman [45], describes the probability distribution for the tree structure describing the ancestral relationships between a present-day sample of haplotypes and their most recent common ancestor (MRCA), together with the associated times for coalescence of each pair of branches and the mutation rate parameter. Kuhner et al. [46] and Griffiths and Tavaré [47, 48] describe MCMC methods for inference in coalescent models. Whereas the coalescent assumes that all the variation between present-day haplotypes is due to mutation, a generalization known as the ancestral recombination graph (ARG) allows for both mutation and recombination, leading to a

graph topology in which joins (moving backwards in time) represent coalescence events and splits represent recombinations [49–51]. Effective MCMC samplers for the ARG have remained elusive, although there has been some progress [50, 52–54]. We are currently exploring forms of rejection sampling known as Approximate Bayesian Computation [55, 56] that avoids the need to compute the likelihood for any given realization of the topology by comparing the ‘closeness’ of random data sampled under that realization to the observed data. For further discussion of coalescent methods for LD mapping, see Zölner [57, this issue].

In a similar vein to our Bayesian spatial clustering model, several authors [58–60] have proposed maximum likelihood methods based on the idea of ancestral haplotype reconstruction from a sample of present-day case and control haplotypes. In these papers, the ancestral haplotype(s) were treated as parameters to be estimated along with the various population parameters (mutation and recombination rates, mutation locations, penetrances, etc.). Morris et al. [61, 62] instead used MCMC methods to allow the ancestral haplotypes to be treated instead as latent variables to be sampled over rather than maximized out.

Haplotype Sharing Methods

An important prediction of coalescent models is that pairs of cases would tend to be more closely related than pairs of controls, while case-control pairs would be even more distantly related on average [63]. This observation underlies a class of LD mapping methods known as haplotype sharing, in which one searches for locations where apparently-unrelated case pairs tend to have more sharing of haplotypes than other pairs. Although this approach has been used informally for meiotic mapping for years, it was first formalized by Te Meerman and Van Der Muelen [64] as a permutation test of a ‘Haplotype Sharing Statistic (HSS),’ compared with a null distribution obtained by randomly permuting the cases and controls. In general, a broad class of HSS can be formulated as $H = \sum_{hk} L_{hk}(x)D_{hk}$ where $L_{hk}(x)$ is the length of the segment surrounding location x that is shared *IBS* by haplotypes h and k , and $D_{hk} = (Y_k - \mu)(Y_h - \mu)$ is a score for the phenotypic similarity of the individuals carrying these haplotypes, the sum being taken over all pairs of haplotypes from apparently unrelated individuals. A number of variants of this general approach have been discussed, including the Haplotype Sharing Correlation [65], the Maxi-

mum Identity Length Contrast (MILC) [66–68], and various other tests [69–75].

In their simplest forms, these various approximations of the coalescent treat each of the present-day haplotypes as independently descended from their respective ancestral haplotype(s); this is known as a ‘star-shaped’ genealogy (or equivalently, the absence of ‘cryptic relatedness’) [76, 77]. McPeck and Strahs and Morris et al. also discuss extensions of their approach to allow for dependency within the sets of descendants from each ancestral haplotype. Our spatial clustering models also have assumed a star-shaped genealogy. One way around this difficulty would be a hierarchical clustering, in which $\gamma_h \sim N(\delta_{c_h}, \sigma^2)$ and $\delta_c \sim N(\eta_{d_c}, \tau^2)$ (where d_c are clusters of clusters), and so on. Reversible jump MCMC methods could be used to allow the number of levels and branches of the hierarchy to be unknown.

A more elegant approach, however, incorporates the estimated time T_{hk} to a common ancestor for each pair of haplotypes within a cluster, based on their shared length L_{hk} , thereby exploiting the notion that haplotypes that are more similar to each other are more likely to be closely related. We therefore allow each haplotype to have its own γ_h with covariance $\text{cov}(\gamma_h, \gamma_k | T_{hk}) = \sigma_c^2 \exp(-\phi T_{hk})$. Since $L_{hk}(x) \sim \Gamma(2, 2T_{hk})$ and $T_{hk} \sim \Gamma(1, 1/(2N))$, where N is the effective population size, the marginal covariance can be shown to be

$$\text{cov}(\gamma_h, \gamma_k | c_h = c_k) = \left(\frac{4 \phi \sigma_c^2}{N} \right) \left(\frac{L_{hk}}{\phi + 1/(2N) + 2L_{hk}} \right)$$

Such methods may provide an alternative approach to the problem of ‘cryptic stratification’ in which a population consists of a number of subpopulations that are not readily distinguishable by self-reported race/ethnicity, but in fact differ both in terms of allele frequencies at the candidate locus and baseline rates of disease (see the reviews by Thomas and Witte [78], Wacholder et al. [79], and Cardon and Palmer [80]). A variety of methods involving the use of a panel of unlinked markers to infer the latent population structure have been proposed [81–84]. We believe such approaches can be readily incorporated into our Bayesian spatial clustering models.

Our implementation of the CAR and Voronoi models currently assume phase-known haplotype data are available, as in a set of transmitted (case) and nontransmitted (control) haplotypes derived from case-parent triads. Extension to the phase-unknown case appears straight-forward, however, simply involving the kinds of Metropolis-Hastings moves discussed above. The acceptance probabilities would now be a function of the relative likelihood

of the cluster assignments of the new and old haplotypes rather than the marginal population haplotype frequencies. Even more appealing is the possibility of exploiting coalescent methods, in which haplotypes might be assigned conditional on the sampled individuals' Y and G together with the current topology of the coalescent tree or ARG, and then the topology and associated parameters are updated conditional on the current assignment of haplotypes.

Haplotype Blocks, Haplotype Tagging SNPs, and the HapMap Project

The recent discovery that the pattern of LD throughout the genome appears to be concentrated in relatively short 'blocks' makes haplotype-based association studies much more attractive. Haplotype blocks are regions with a limited number of distinct haplotypes separated by regions of low LD. The latter could reflect either hot spots of recombination or ancient recombination events [85–87]. Most importantly, it implies that to identify all the polymorphisms within a region that are relevant to disease, one could simply select a subset of 'haplotype-tagging SNPs' (htSNPs) that in combination are sufficient to predict the all the common haplotypes within each block.

A number of approaches to htSNP selection have been described [8, 9, 88–94], but the number of htSNPs which are needed in whole-genome association studies is not presently known. There is currently underway a massive resequencing effort, known as the HapMap Project [95, 96] (see also NIH News Advisory, October 2002, at <http://www.nih.gov/news/pr/oct2002/nhgri-29.htm>), which is aimed at characterizing haplotype blocks across the entire genome for the purpose of selecting htSNPs for a new generation of association studies. The hope is that this effort will reduce the number of SNPs that would be required for genome-wide association studies (as suggested by Risch and Merikangas [97]) from several million by perhaps an order of magnitude.

There is some controversy, however, about whether the number of htSNPs needed may actually be reduced to this extent. Very recently, Carlson et al. [98] compared the results of their own SNP discovery (sequencing) efforts with SNPs already deposited in the dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP/>) and estimated (by extrapolation from their results for 50 genes) that all 2.7 million known SNPs in that database are barely sufficient in whites and quite insufficient in African Americans to accurately predict the majority of their newly discovered

SNPs. Unfortunately, the statistic they used to determine whether a given SNP discovered by their sequencing was well predicted by the dbSNP database was the maximum of simple pairwise R^2 s (i.e., for each new SNP discovered in the sequencing, the maximal pairwise R^2 between that SNP and each dbSNP was computed). Simple pairwise R^2 may be a very poor measure of the actual predictability of either common SNPs or (which may be of even more importance) of common haplotypes based on any given set of htSNPs.

Our approach [99] for determining the number and identity of htSNPs is based on maximizing the minimum across all common haplotypes of the statistic R_h^2 , the expected correlation between the true number $N_h(G_i)$ of copies of haplotype h for each individual and its predicted value given the full subset of SNPs. This approach fully exploits the multivariate pattern of LD but requires reasonably good estimation of all the haplotype frequencies in a particular block (and is dependent on H-W equilibrium). An initial simulation [99] found good behavior of the R_h^2 statistic when the true state of nature was that of quite limited haplotype diversity within a block. In this simulation, however, 70 individuals of a particular ethnic group were used to characterize the haplotypes, which is considerably more than the number of subjects sequenced by Carlson et al. (23 African Americans and 24 Whites). Thompson et al. [100, this issue] do consider preliminary studies as small as this of the selection of htSNPs, with promising results found when using a coalescent-based simulation of both haplotypes and disease genes. A computer program to compute R_h^2 and find minimal sets of htSNPs using this criterion is available from our website (<http://www-rcf.usc.edu/~stram/tagsnpsv1.zip>).

Chapman et al. [101, this issue] describe an alternative approach based on choosing the subset of SNPs that maximizes the minimum across loci l of R_l^2 , the multiple correlation between G_{il} and the subset, on the assumption that the causal variant being sought is likely to be in the dataset; this measure again is distinguished from the pairwise R_l^2 used by Carlson et al. [98] in that it exploits the multivariate pattern of LD. If a set of haplotype frequency estimates are available (e.g. by running an E-M algorithm), a formal calculation of the expected (squared) correlation R_l^2 can be performed [99], rather than estimating this squared correlation directly from a regression analysis of non-htSNP genotypes upon htSNP genotypes. This formal calculation is dependent upon the assumption of H-W equilibrium (for the haplotypes). It is possible that the formal calculation is a more efficient estimate of R_l^2 if H-W equilibrium does indeed hold.

Even if htSNP characterization is successful in reducing the total number of variants needed to be typed in association studies, it remains to be seen whether such common variants are really the cause of common complex diseases [102]. The computational challenges posed by trying to identify haplotype blocks and haplotypes from genome-wide diploid genotype data are formidable and

an area deserving further attention. We are continuing to pursue these issues from a Bayesian perspective, which we believe offers the prospect of providing a unified framework that adequately allows for the uncertainty in our characterization of the block structure itself when drawing inferences about the best subset of htSNPs and their association with disease.

References

- 1 Botstein D, Risch N: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;33:228–237.
- 2 Sachidanandam R, Weissman D, Schmidt SC, et al: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–933.
- 3 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;12:921–927.
- 4 Chiano M, Clayton D: Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet* 1998;62:55–60.
- 5 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;70:425–434.
- 6 Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG: Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 2002;53:79–91.
- 7 Stram DO, Pearce CL, Bretsky P, et al: Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered* 2003;55:180–191.
- 8 Daly M, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001;29:229–232.
- 9 Patil N, Berno AJ, Hinds DA, et al: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001;294:1719–1723.
- 10 Niu T, Qin ZS, Xu X, Liu JS: Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002;70:157–169.
- 11 Liu JS, Sabatti C, Teng J, Keats BJB, Risch N: Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res* 2001;11:1716–1724.
- 12 Lin S, Cutler DJ, Zwick ME, Chakravarti A: Haplotype inference in random population samples. *Am J Hum Genet* 2002;71:1129–1137.
- 13 Zhao LP, Li SS, Khalid N: A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet* 2003;72:1231–1250.
- 14 Akey J, Jin L, Xiong M: Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* 2001;9:291–300.
- 15 Bader JS: The relative power of SNPs and haplotype as genetic markers for association tests [comment]. *Pharmacogenomics* 2001;2:11–24.
- 16 Morris RW, Kaplan NL: On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002;23:221–233.
- 17 Long AD, Langley CH: The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999;9:720–731.
- 18 Judson R, Stephens JC, Windemuth A: The predictive power of haplotypes in clinical response. *Pharmacogenomics* 2000;1:15–26.
- 19 Conti DV, Witte JS: Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet* 2003;72:351–363.
- 20 Madigan D, Raftery AE: Model selection and accounting for model uncertainty in graphical models using Occam's window. *J Am Statist Assoc* 1994;89:1335–1346.
- 21 Raftery AE, Madigan D, Hoeting JA: Bayesian model averaging for linear regression models. *J Am Statist Assoc* 1997;92:179–191.
- 22 Conti DV, Cortessis V, Molitor J, Thomas DC: Bayesian modeling of complex metabolic pathways. *Hum Hered* 2003;56:83–93.
- 23 Anderson JA: Separate sample logistic discrimination. *Biometrika* 1972;59:19–35.
- 24 Prentice R, Breslow N: Retrospective studies and failure time models. *Biometrika* 1978;65:153–158.
- 25 Whittemore A: Logistic regression of family data from case-control studies. *Biometrika* 1995;82:57–67.
- 26 Whittemore AS, Halpern J: Logistic regression of family data from retrospective study designs. *Genet Epidemiol* 2003, submitted.
- 27 Greenland S, Morgenstern H, Thomas DC: Considerations in determining matching criteria and stratum sizes for case-control studies. *Int J Epidemiol* 1981;10:389–392.
- 28 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003, in press.
- 29 Kraft P, Thomas DC: Bias and efficiency in family-matched gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. *Am J Hum Genet* 2000;66:1119–1131.
- 30 Clayton D, Kaldor J: Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987;43:671–681.
- 31 Mollie A, Richardson S: Empirical Bayes estimates of cancer mortality rates using spatial models. *Stat Med* 1991;10:95–112.
- 32 Knorr-Held L, Rasser G: Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* 2000;56:13–21.
- 33 Denison DG, Holmes CC: Bayesian partitioning for estimating disease risk. *Biometrics* 2001;57:143–149.
- 34 Thomas DC: Bayesian models and Markov chain Monte Carlo methods. *Genet Epidemiol* 2001;21(suppl 1):S660–S661.
- 35 Molitor J, Marjoram P, Thomas D: Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. *Genet Epidemiol* 2003, in press.
- 36 Green P, Richardson S: Hidden Markov models and disease mapping. *J Am Stat Assoc* 2002;97:1055–1070.
- 37 Voronoi MG: Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J Reine Angew Math* 1908;34:198–287.
- 38 Molitor J, Marjoram P, Thomas DC: Application of Bayesian clustering via Voronoi tessellations to the analysis of haplotype risk and gene mapping. *Am J Hum Genet* 2003, submitted.
- 39 Green P: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *J R Statist Soc, Ser B* 1995;82:711–732.
- 40 Hill W, Weir B: Maximum-likelihood estimation of gene location by linkage disequilibrium. *Am J Hum Genet* 1994;54:705–714.
- 41 Kaplan N, Hill W, Weir B: Likelihood methods for locating disease genes in nonequilibrium populations. *Am J Hum Genet* 1995;56:18–32.

- 42 Xiong M, Guo SW: Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. *Am J Hum Genet* 1997;60:1513–1531.
- 43 Graham J, Thompson E: Disequilibrium likelihoods for fine-scale mapping of a rare allele. *Am J Hum Genet* 1998;63:1517–1530.
- 44 Zollner S, von-Haeseler A: A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet* 2000;66:615–628.
- 45 Kingman JFC: On the genealogy of large populations. *J Appl Prob* 1982;19A:27–43.
- 46 Kuhner M, Yamato J, Felsenstein J: Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 1995;140:1421–1430.
- 47 Griffiths R, Tavaré S: Simulating probability distributions in the coalescent. *Theor Pop Biol* 1994;46:131–159.
- 48 Griffiths R, Tavaré S: Computational methods for the coalescent; in Donnelly P, Tavaré S (eds): *Population Genetics and Human Evolution*. Berlin, Springer, 1997, pp 165–182.
- 49 Hudson R: Properties of a neutral allele model with intragenic recombination. *Theor Pop Biol* 1983;23:183–201.
- 50 Griffiths R, Marjoram P: Ancestral inference from samples of DNA sequences with recombination. *J Comp Biol* 1996;3:479–502.
- 51 Griffiths R, Marjoram P: An ancestral recombination graph; in Donnelly P, Tavaré S (eds): *Progress in Population Genetics and Human Evolution/IMA Volumes in Mathematics and Its Applications*. Berlin, Springer, 1997, pp 257–270.
- 52 Kuhner M, Felsenstein J: Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genet Epidemiol* 2000;19:S15–S21.
- 53 Fearnhead P, Donnelly P: Estimating recombination rates from population genetic data. *Genetics* 2001;159:1299–1318.
- 54 Larribe F, Lessard S, Schork NJ: Gene mapping via the ancestral recombination graph. *Theor Pop Biol* 2002;62:215–229.
- 55 Beaumont MA, Zhang W, Balding DJ: Approximate Bayesian Computation in Population Genetics. *Genetics* 2002;162:2025–2035.
- 56 Marjoram P, Molitor J, Plagnol V, Tavaré S: Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 2003, in press.
- 57 Zöllner S, Pritchard J: Mapping genes with the use of a local approximation to the ancestral recombination graph (abstract). *Hum Hered* 2003;56:149.
- 58 McPeck M, Strahs A: Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. *Am J Hum Genet* 1999;65:858–875.
- 59 Service S: Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Am J Hum Genet* 1999;64:1728–1738.
- 60 Seltman H, Roeder K, Devlin B: Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 2001;68:1250–1263.
- 61 Morris A, Whittaker J, Balding D: Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am J Hum Genet* 2000;67:155–169.
- 62 Morris AP, Whittaker JC, Balding DJ: Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet* 2002;70:686–707.
- 63 Markovtsova L, Marjoram P, Tavaré S: The age of a unique event polymorphism. *Genetics* 2000;156:401–409.
- 64 Te Meerman G, Van Der Meulen M: Genomic sharing surrounding alleles identical by descent effects of genetic drift and population growth. *Genet Epidemiol* 1997;14:1125–1130.
- 65 Qian D, Thomas D: Genome scan of complex traits by haplotype sharing correlation. *Genet Epidemiol* 2001;21:S582–S587.
- 66 Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* 2000;64:255–265.
- 67 Bourgain C, Genin E, Holopainen P, et al: Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *Am J Hum Genet* 2001;68:154–159.
- 68 Bourgain C, Genin E, Margaritte-Jeannin P, Clerget-Darpoux F: Maximum identity length contrast: a powerful method for susceptibility gene detection in isolated populations. *Genet Epidemiol* 2001;21:S560–S564.
- 69 Boon M, Nolte IM, Bruinenberg M, et al: Mapping of a susceptibility gene for multiple sclerosis to the 51 kb interval between G511525 and D6S1666 using a new method of haplotype sharing analysis. *Neurogenetics* 2001;3:221–230.
- 70 Devlin B, Roeder K, Wasserman L: Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing. *Biostatistics* 2000;1:369–387.
- 71 Fischer C, Beckmann L, Majoram P, Te Meerman G, Chang-Claude J: Haplotype sharing analysis with SNPs in candidate genes: the Genetic Analysis Workshop 12 example. *Genet Epidemiol* 2003;24:68–73.
- 72 Nolte IM: The Haplotype Sharing Statistic: Fine-mapping of disease gene loci by comparing patients and controls for the length of haplotype sharing. Groningen: University of Groningen, 2002.
- 73 Schork N, Thiel B, St. Jean P: Linkage analysis, kinship, and the short-term evolution of chromosomes. *J Exp Zool* 1999;282:133–149.
- 74 Tzeng JY, Devlin B, Wasserman L, Roeder K: On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 2003;72:891–902.
- 75 Zhang S, Zhao H: Linkage disequilibrium mapping in population of variable size using the decay of haplotype sharing and a stepwise-mutation model. *Genet Epidemiol* 2000;19: S99–S105.
- 76 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- 77 Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. *Theor Pop Biol* 2001;60:226–237.
- 78 Thomas DC, Witte JS: Point: Population stratification: A problem for case-control studies of candidate gene associations? *Cancer Epidemiol Biomarkers Prev* 2002;11:505–512.
- 79 Wacholder S, Rothman N, Caporaso N: Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiologic studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* 2002;11:513–520.
- 80 Cardon LR, Palmer LJ: Population stratification and spurious allelic association. *Lancet* 2003;361:598–604.
- 81 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000;67:170–181.
- 82 Satten GA, Flanders WD, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;68:466–477.
- 83 Devlin B, Roeder K, Bacanu S-A: Unbiased methods for population-based association studies. *Genet Epidemiol* 2001;21:273–284.
- 84 Reich DE, Goldstein DB: Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 2001;20:4–16.
- 85 Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 2002;71:1227–1234.
- 86 Zhang K, Akey JM, Wang N, Xiong M, Chakraborty R, Jin L: Randomly distributed crossovers may generate block-like patterns of linkage disequilibrium: An act of genetic drift. *Hum Genet* 2003;3:3.
- 87 Arnheim N, Calabrese P, Nordborg M: Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet* 2003;73:5–16.
- 88 Clayton DG: Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. Unpublished manuscript, cited in Johnson CG, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;29:233–237. Available from <http://www.nature.com/ng/journal/v29/n2/extref/ng1001-233-S10.pdf>, 2002.
- 89 Mannila H, Koivisto M, Perola M, et al: Minimum Description Length Block Finder, a method to identify haplotype blocks and to compare the strength of block boundaries. *Am J Hum Genet* 2003;73:86–94.
- 90 Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet* 2003;73:115–130.

- 91 Qin ZS, Niu T, Liu JS: Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet* 2002;71:1242–1247.
- 92 Zhang K, Calabrese P, Nordborg M, Sun F: Haplotype block structure and its applications to association studies: Power and study designs. *Am J Hum Genet* 2002;71:1386–1394.
- 93 Zhang K, Deng M, Chen T, Waterman MS, Sun F: A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* 2002;99:7335–7339.
- 94 Zhang K, Sun F, Waterman MS, Chen T: Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am J Hum Genet* 2003;73:63–73.
- 95 Cardon LR, Abecasis GR: Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003;19:135–140.
- 96 Couzin J: Genomics. New mapping project splits the community. *Science* 2002;296:1391–1393.
- 97 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996;273:1616–1617.
- 98 Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA: Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* 2003;33:518–521.
- 99 Stram DO, Haiman CA, Hirschhorn JN, et al: Choosing haplotype-tagging SNPs based on unphased genotype data from a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum Hered* 2003;55:27–36.
- 100 Thompson D, Stram D, Goldgar D, Witte JS: Haplotype tagging single nucleotide polymorphisms and association studies. *Hum Hered* 2003;56:48–55.
- 101 Chapman JM, Cooper JD, Todd JA, Clayton DG: Detecting disease associations due to linkage disequilibrium using haplotype tags: A class of tests and the determinants of statistical power. *Hum Hered* 2003;56:18–31.
- 102 Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;69:124–137.